Ethnicity Effects in Police Officer Selection

Applicant, Assessor, and Selection-Method Factors

Lonneke de Meijer

The research presented in this dissertation was supported in part by the Police Academy of The Netherlands and the Dutch Foundation for Psycho Technique (Nederlandse Stichting voor Psychotechniek; NSvP).

ISBN 978-90-76269-69-6

Cover designed by Ridderkerk B.V., Ridderkerk
Painting on cover by Sarah Solie "Diversity"
Lay-out by Lonneke A. L. de Meijer
Printed by Ridderkerk B.V., Ridderkerk

Ethnicity Effects in Police Officer Selection: Applicant, Assessor, and Selection-Method Factors

Effecten van etniciteit bij de selectie van politieagenten: kandidaats-, beoordelaar-, en selectiemethodefactoren

Proefschrift

ter verkrijging van de graad van doctor aan de Erasmus Universiteit Rotterdam op gezag van de rector magnificus

Prof.dr. S.W.J. Lamberts

en volgens besluit van het College voor Promoties. De openbare verdediging zal plaatsvinden op

vrijdag 19 september 2008 om 13:30uur

door

Lonneke Adriana Lucia de Meijer
geboren te Wagenberg

Promotiecommissie

Promotoren:	Prof.dr. H.T. van der Molen Prof.dr. M.Ph. Born
Overige leden:	Prof.dr. H.G. Schmidt Prof.dr. J. von Grumbkow Prof.dr. F.J.R. van de Vijver

Contents

Chapter 1

Introduction: Ethnicity effects in personnel selection

A field in which differences on psychological measures between ethnic majority and minority groups have been extensively investigated is the domain of personnel selection. Most of the research on ethnic score differences has been conducted in North America and has treated ethnic minorities as one homogeneous group, which merely has been contrasted with the ethnic majority group. That is, a dichotomous distinction has been made between Whites and non-Whites or between the majority and the minority group. This approach ignores the many visible and cultural differences among ethnic minority groups, both in the U.S. and outside, that may affect scores on selection instruments. In an attempt to overcome this limitation, the present dissertation examined the largest ethnic minority groups in The Netherlands, i.e., Dutch Antillean, Moroccan, Surinamese, and Turkish ethnic minority groups. The focus was to obtain a more detailed picture of the differences between the various ethnic groups in Dutch society. A second limitation of the existing literature is its descriptive character. Attempts to present possible explanations for the existing differences between ethnic groups have hardly been given. To fill this gap, five empirical studies have been conducted to examine applicant, assessor, and selection-method factors, which potentially are related to score differences between ethnic groups on a series of tests. These are a cognitive ability test, a personality questionnaire, an assessment center, an employment interview, a final employment recommendation, and a situational judgment test. The general project goal was to map the relative extent to which these factors are able to explain existing ethnic score differences.

The Netherlands is a country that consists of a population of 16.3 million people of which more than 19% are ethnic minorities (both Western and non-Western ethnic minority group members; CBS, January 1, 2007). Several definitions of ethnic minority versus majority people exist. In the present dissertation, the following commonly used definitions are utilized. An ethnic minority person is someone who is born outside The Netherlands or someone whose parents (or at least one of the parents) are born outside The Netherlands. An ethnic majority person is someone who is born in The Netherlands and whose parents are born in The Netherlands (CBS, January 1, 2007).

People from four large non-Western countries inhabit The Netherlands, which are the largest ethnic minority groups in The Netherlands, namely 129,965 from the Dutch Antilles (and Aruba), 329,493 from Morocco, 333,504 from Surinam, and 368,600 from Turkey. Together, the people from these four non-Western ethnic minority groups form 7% of the Dutch population. In 2006, a working population of 68% existed in The Netherlands (compared to the total population between 15 and 64 years old). The ethnic majority labor force equaled 70% compared to 67% of the Dutch Antilleans, 47% of the Moroccans, 68% of the Surinamese, and 52% of the Turks (CBS, January 1, 2007). The unemployment rate in The Netherlands was 6% in 2006. The ethnic majority unemployment was 4% compared to 12% of the Dutch Antilleans, 17% of the Moroccans and the Surinamese, and 15% of the Turks (CBS, January 1, 2007). These numbers exhibit a relatively low presence of ethnic minority group members in the Dutch working population compared to the ethnic majority group. Less ethnic minority people entering the labor market or more ethnic minority people leaving the labor market could cause this unequal distribution of ethnic groups in the working population. The present dissertation will shed more light on factors that may influence hiring opportunities of ethnic majority and minority group members during personnel selection. First, an overview will be given of the migration history of ethnic minority groups into The Netherlands. Second, a more detailed picture will be drawn of ethnicity issues in personnel selection in the U.S. and Europe. Finally, applicant, assessor, and method factors of potential influence on personnel-selection decisions will be outlined, with each factor resulting in several research questions.

1.1 History of Ethnic Groups in The Netherlands

Although migration to The Netherlands existed for centuries before 1900, the number of immigrants increased continuously in the beginning of the 20th century. Migration mainly started between WWI and WWII from China, the Dutch Antilles, and the Dutch colonies (the Dutch Indies, Surinam) and increased massively after WWII because of decolonization (Hoving, Dibbits, & Schrover, 2005; Vogel, 2005). After 1960, a shift occurred from colonial migration to the arrival of 'migrant workers' from Mediterranean countries. Their specific cultural influence increased over the years, specifically when they stayed permanently (Hoving et al., 2005; Vogel, 2005). The migration of people from the Mediterranean partly coincided with the migration from Surinam after its independence in 1975. Migration of people seeking political asylum originated only after 1975 (Hoving et al., 2005; Vogel, 2005). The period from 1950 onwards will now be described in more detail.

During the 1950s, more and more non- or semi-skilled – predominantly male – workers were required in The Netherlands. Especially from 1960 onwards, Dutch government and businesses started actively enlisting migrant workers from the Mediterranean, especially from Italy at first. When recruitment from Italy fell short, recruitment was extended to other Mediterranean countries such as Spain, Greece, and – more importantly – Turkey and Morocco (Lucassen, 2005; Lucassen & Penninx, 1994). In 1973, the recruitment of migrant workers was stopped because of an economic recession in The Netherlands. Contrary to expectations, however, the number of migrants increased steadily until today. This increase was caused by a combination of poor economic perspectives in the Mediterranean and, paradoxically, a restrictive Dutch government policy against migrants. The latter factor made migrant workers, especially from Turkey and Morocco, realize that if they would leave The Netherlands they would not be allowed back into this country. Furthermore, most of them still were employed and had become used to living in The Netherlands. Moreover, over the years they had become entitled to reunification with their family. Instead of leaving The Netherlands, therefore, they had their wives, children, and families come over to The Netherlands. This caused an influx of migrants that was much larger than the original number of immigrants from the 1950s and 1960s (Lucassen, 2005; Lucassen & Penninx, 1994).

In the beginning of the 1970s, a large number of colonial and non-colonial migrants started to migrate to The Netherlands. Migration occurred especially from Surinam, as Surinam became independent from The Netherlands in

1975, and also from the Dutch Antilles. During the entire 20th century, Surinamese people migrated to The Netherlands. At the beginning of this century, this was a relatively small group of predominantly students, highly educated people, musicians, migrant workers, and nurses. However, when in sight of the independence of 1975, a conflict between different sections of the Surinamese population became more and more visible and The Netherlands appeared to have more than enough jobs, a third of the Surinamese population decided to move to The Netherlands to look for good fortune (Lucassen, 2005; Lucassen & Penninx, 1994). Also around this time, a relatively large amount of people from the Dutch Antilles came to The Netherlands to search for a better life (although smaller in number than the numbers of Moroccans, Surinamese, and Turks).

1.2 Personnel Selection in a Multi-Ethnic Setting

The unequal representation of the four largest ethnic minority groups in the Dutch working population compared to the ethnic majority group may be caused by different reasons, among which personnel selection strategies and turnover at work. The present dissertation focuses on explanations for this unequal representation in the selection process. Differences on selection measures between ethnic majority and ethnic minority groups have been extensively investigated, with a particular focus on cognitive ability (or *g*), as *g* has been found to be a consistently good predictor of job performance across a variety of occupations (Schmidt & Hunter, 1998, 2004). In particular for more complex job levels, the predictive validity of *g* is high (Hunter, 1986). At the same time, however, several studies (e.g., Goldstein, Zedeck, & Goldstein, 2002; Murphy, 2002; Outtz, 2002) have shown that the cognitive ability test represents the predictor that most likely will have substantial adverse impact on employment opportunities for most ethnic minority groups.

When employers want to maximize the skill level of their employees on the one hand and diversify their workforce on the other hand, both goals cannot be achieved at the same time because of existing subgroup differences on the cognitive ability test. A possible solution for this dilemma has been sought in the use of other (non-cognitive ability) selection measures, e.g., the assessment center (AC), the employment interview, and the situational judgment test (SJT). These measures have the advantage of showing smaller score differences between ethnic majority and minority groups and, consequently, a lower adverse impact on employment opportunities than cognitive measures (Murphy, 2002). Little is known, however, about the

possible factors that may influence score differences between ethnic groups on cognitive and non-cognitive selection measures. The purpose of the present dissertation is to shed more light on applicant, assessor, and method-related factors explaining score differences between the ethnic majority group and the four largest ethnic minority groups in The Netherlands, i.e., Dutch Antillean, Moroccan, Surinamese, and Turkish ethnic minority groups. To this end, we conducted five empirical studies at the Dutch police.

The Dutch police is an interesting organization in terms of its personnel selection procedure and the heterogeneity of its employees. First, since 1985, the Dutch police has aimed at a percentage of ethnic minority employees that would be a reflection of the percentage of ethnic minorities in Dutch society. Later on, the Dutch police set a more concrete goal: In 2002, 10% of the work force should be from non-Western descent (Broekhuizen, Raven, & Driessen, 2007). In 2005, the percentage of (non-Western) ethnic minority employees at the Dutch police was 6% (LECD, 2006), showing that the intended goal of 10% had not been achieved yet, despite numerous changes in recruitment, selection, employment, and career planning. Second, several reasons for a multi-ethnic police force have been put forward over the years. At first, the Dutch police wanted to be an example for multi-cultural public institutions and therefore wanted its workforce to be a reflection of Dutch society (Broekhuizen et al., 2007). The aim of being an example for other organizations resulted in the ambition to have 10% ethnic minority employees. Besides being an example for other organizations concerning different ethnic groups and their job opportunities, employing ethnic minority employees turned out to be advantageous for the police in several ways, e.g., having knowledge of the language, culture and religion of ethnic minority groups, addressing and contacting ethnic minority people, and being a model for ethnic minority youth (Broekhuizen et al., 2007). The following situation reflects some of the benefits of employing ethnic minority police(wo)men:

> "When I walked through the door, I could instantly see that the man was reassured, because he would be understood. Immediately, the atmosphere was based more on trust than would have been the case when an ethnic majority policeman would be involved. By means of having a conversation about what the Koran teaches, in the end I reached some sort of compromise where I asked this man to respect the Dutch rules. Things were acknowledged sooner than would have been the case with an ethnic majority policeman" (Broekhuizen et al., 2007, p. 54).

Despite the positive influence ethnic minority employees may have on the Dutch police organization, obviously, their potential is not yet fully utilized

(LECD, 2006). Among the reasons for this are recruitment strategies that do not reach ethnic minority target groups, relatively high ethnic minority dropouts during police training, career perspectives that are not as promising for ethnic minority employees as they are for ethnic majority employees, and personnel selection and hiring opportunities that are unequal for different (ethnic) groups. It is the latter issue that is the focus of this dissertation. The following potentially explanatory factors are distinguished when investigating the differences between the ethnic majority group and the ethnic minority groups: 1) individual differences among applicants, 2) assessor-related differences, and 3) the context (selection method) within which the selection process takes place (cf. Klimoski & Donahue, 2001). Our main research question, therefore, is:

> To what extent do applicant, assessor, and method-related factors explain score differences on selection measures between the ethnic majority group and the ethnic minority groups?

Before testing the three factors, however, the existing score differences between the ethnic majority group and the ethnic minority groups are investigated, as it is useful only to examine possible explaining factors on score differences when score differences actually exist. Tot this end, the first part of Chapter 2 serves as a starting point for further research on the topic of possible explanatory factors of score differences on selection measures. Selection measures can differ in the extent of assessor influence on the scores. When there is no assessor influence (e.g., cognitive ability tests and personality questionnaires), the selection measure is labeled as an *objective measure*. By contrast, measures in which an assessor is involved (e.g., assessment centers [ACs] and employment interviews) are labeled *subjective measures* (Bass & Barrett, 1981). In Chapter 2, score differences on the objective measures (i.e., the cognitive ability test and the personality questionnaire) and on the subjective measures (i.e., the AC, the employment interview, and the final employment recommendation) are calculated, compared with each other, and compared to score differences on these measures in North America and Europe. Furthermore, the four largest ethnic minority groups in The Netherlands each are compared to the Dutch majority group separately, to examine to what extent the minority groups differ from the majority group. In addition, first-generation minority groups (i.e., individuals born outside The Netherlands) are compared to second-generation minority groups (i.e., individuals born in The Netherlands, but al least one of the parents is born outside The Netherlands) to display which groups improve most and least from one generation to another. Consequently the following research question is formulated.

Research question 1: What are the score differences on the various objective and subjective selection measures?

a) What are the score differences for the various ethnic minority groups compared to the ethnic majority group on the selection measures?

b) What are the score differences between first-generation-minority groups and the second-generation minority groups on the selection measures?

The applicant, assessor, and selection-method factors that are investigated in the five empirical studies of the present dissertation will now be discussed in more detail.

Applicant Factors

The second part of Chapter 2 focuses on how ethnicity-related applicant factors, such as language-proficiency and education, may influence test scores on different types of selection measures. First, concerning the objective measures (i.e., the cognitive ability test and the personality questionnaire), the explanatory power on score differences between ethnic groups is investigated of the following applicant demographics: (a) Dutch language-proficiency, (b) education, and (c) ethnicity. Second, concerning the subjective measures (i.e., the AC, the employment interview, and the final employment recommendation) two theoretical perspectives developed within social psychology are taken out of the laboratory and tested in a field setting. These perspectives are assumed-characteristics theory (Locksley, Borgida, Brekke, & Hepburn, 1980; Locksley, Hepburn, & Ortiz, 1982a; 1982b) and complexity-extremity theory (Linville, 1982; Linville & Jones, 1980). These theories can be described as follows.

Assumed-characteristics theory suggests that, based on knowledge about certain demographics of a group, people make assumptions about other characteristics of this group. For example, on the basis of knowledge of a group's ethnicity, people assume that this group will have a certain socio-economic status (SES), education, or personality. Assumed-characteristics theory suggests that members of an in-group will believe their own characteristics are more favorable than characteristics of members of an out-group (Coleman, Jussim, & Kelley, 1995; Jussim, Coleman, & Lerch, 1987; Jussim, Fleming, Coleman, & Kohberger, 1996). Having more relevant ethnicity-related demographic information about out-group members should nevertheless substantially decrease the unfavorable assumed characteristics and evaluations of out-group members should become more positive.

Complexity-extremity theory (Linville, 1982; Linville & Jones, 1980) starts with the assumption that people have more contact with in-group members than with out-group members. Because in-group members have more contact with other in-group members, they will develop more complex representations of in-group members than of out-group members. When observers use these more complex representations to evaluate an in-group member, they are likely to give accurate evaluations. Low complex or simple representations will be developed about out-group members. When these simple representations are used in evaluating an out-group member, extreme evaluations are more likely because the out-group member can more easily be seen as all good or all bad.

In testing these theories, it will be endeavored to make a distinction among ethnic minority groups, as until now these has often been a tendency in existing research to lump ethnic minorities together in one group and to contrast this group with the ethnic majority group. A more differentiated perspective is needed when different ethnic groups are involved. In sum, the second research question thus relates to applicant factors that may explain score differences on various selection measures.

> *Research question 2:* Are applicant factors able to explain score differences between ethnic groups on personnel selection measures?
>
> a) Are applicant demographics, such as language proficiency, education, and ethnicity, able to explain score differences between ethnic groups on objective measures (i.e., the cognitive ability test and the personality questionnaire) and subjective measures (i.e., the AC, the employment interview, and the final employment recommendation)?
>
> b) Are ethnic minority applicants whose demographics are either very positive or very negative evaluated more extremely (i.e., positive or negative, respectively) than ethnic majority applicants with the same demographics?

Assessor Factors

In Chapters 3 and 4, assessor factors and their explanatory power regarding differences between ethnic groups of applicants on subjective measures (cf. Bass & Barrett, 1981) are investigated. The purpose of Chapter 3 is to examine differences that may exist between the judgment processes of ethnic majority assessors judging ethnic majority applicants and of ethnic majority assessors judging ethnic minority applicants. With the term 'judgment process' we mean the process of giving weights to sources of information (e.g., scores on a personality questionnaire and an AC exercise) when combining these into a final employment recommendation.

Judgment analysis is a methodological application of Social Judgment Theory (SJT) and its underlying framework, Brunswick's Lens Model (e.g., Brunswick, 1952). SJT defines judgment as a process that involves the integration of information from a set of cues into a judgment about certain outcomes (e.g., selection results/outcomes, in case of personnel selection). Usually, for judgment analysis, a statistical model is defined for a specific assessor by means of multiple regression analysis. The resulting regression equation represents the strategy of the assessor and the regression weights reflect the importance of certain variables or cues awarded by this assessor. Judgment analysis mainly consists of analyses that allow the identification of the weights assigned to pieces of available information during decision-making.

To our knowledge, the effect of applicant ethnicity on the assessor's judgment process has not been investigated until now. Pulakos, White, Oppler, and Borman (1989) argued that irrespective of whether there are mean subgroup differences in judgments of assessors, assessors may use different variables or cues in their process of judging someone of a different ethnical background. Judgment analysis is a possible strategy for investigating such similarities and differences in judgment processes. The following research question can be derived from the area of judgment analysis.

> *Research question 3:* Do assessors integrate information into a final employment recommendation in a different way when judging an ethnic majority applicant than when judging an ethnic minority applicant?

Chapter 4 also focuses on assessors and their effect on subjective measures, but takes a different perspective than Chapter 3. During interpersonal perception many factors may influence impressions and inferences made by an assessor, among which affective processes, interpersonal factors, and motivation and skills of the assessor. With regard to interpersonal factors, the similarity between the assessor and the assessee may be expected to have an influence on the outcome of perceptual processes (Fiske & Taylor, 1991; Klimoski & Donahue, 2001). It is this similarity issue, which is the focus of Chapter 4. More specifically, it is investigated to what extent demographic, in this case ethnic, similarity and perceived similarity between assessors and applicants are able to explain existing score differences between ethnic groups on several subjective instruments. These are the AC, the employment interview, and the final employment recommendation.

Again, two social psychological theories are taken out of the laboratory and tested in the field of personnel selection. Regarding demographic similarity,

Social Identity Theory (SIT; Tajfel, 1982; Turner, 1987) is tested, which contends that aspects of an individual's self image come from the social categories to which he/she perceives him/herself as belonging, such as ethnic group and gender. Social identity is seen as necessary to boost one's self esteem. To the extent that individuals' social identities and self-categorizations are built around their demographic characteristics, demographic dissimilarity may have a negative effect on the attitudes and behaviors towards others, whereas higher identification and similarity may lead to more positive attitudes and behaviors towards other people. Concerning perceived similarity, perceived intergroup similarity is examined. Here, the focus is not on similarity in terms of objective demographic characteristics but on perceptions of similarity, including less tangible attributes such as values, beliefs, and personality.

Because of the nested data structure of the study presented in Chapter 4, the more appropriate multilevel analysis technique is used to examine the effects of both demographic and perceived similarity on score differences between ethnic groups. The question rises whether the effects regarding demographic and perceived similarity that often are reported in the literature based on classic ANOVA are truthful reflections of reality. The same question rises concerning the social psychological theories that have found support in laboratory settings, but may not last in a field setting. The following research question has been formulated related to demographic and perceived similarity.

> *Research question 4:* Do demographic and perceived similarity between assessor and applicant explain score differences between the ethnic majority group and ethnic minority groups?

Method Factors

The studies presented in Chapters 2 to 4, applicant and assessor factors are examined. Method factors are examined in Chapters 5 and 6. With method factors, we mean the context in which selection takes place. More specifically, in Chapters 5 and 6, it is investigated whether method factors, i.e., the psychological measures that are used, explain differences between the ethnic majority group and ethnic minority groups.

Among the selection measures, cognitive ability (or *g*) has been consistently found to be the best predictor of job performance across a variety of occupations (Schmidt & Hunter, 1998, 2004). The question rises, whether the skills of all employees are indeed maximized when general mental ability (or *g*) is used as a selection measures? Regarding this question, several studies (e.g.,

Goldstein, Zedeck, & Goldstein, 2002; Murphy, 2002; Outtz, 2002) have shown that the cognitive ability test represents the predictor that most likely will have substantial adverse impact on employment opportunities for most ethnic minority groups. A possible solution for this dilemma has been sought in the use of other (non-cognitive ability) selection measures, e.g., the AC and the employment interview. These have the advantage of showing smaller score differences between ethnic groups and, consequently, a lower adverse impact on employment opportunities than cognitive measures (Murphy, 2002). In Chapter 5, the predictive power of cognitive and non-cognitive ability selection measures is investigated. In addition, differential prediction of these measures is explored. The following research question has been formulated related to (differential) predictive validity.

> *Research question 5:* Is the predictive validity of the non-cognitive ability tests comparable to the predictive validity of the cognitive ability test for both the ethnic majority and the ethnic minority group?

Concerning the investigation into the effect of method factors on score differences between ethnic groups, Chapter 6 describes a type of test that has become more and more popular in the last two decades, namely the Situational Judgment Test (SJT). The SJT refers to a test that typically consists of hypothetical scenarios describing a work situation in which a problem has arisen and has been advocated as a means of diminishing score differences between ethnic groups (e.g., Weekley & Jones, 1997, 1999). Two types of SJT formats exist namely the paper-and-pencil SJT and the video-based SJT. Compared to paper-and-pencil SJTs, video-based SJTs appear to have the additional advantages of showing higher predictive validity (Lievens & Sackett, 2006) and smaller score differences between ethnic groups (Chan & Schmitt, 1997; Lievens & Sackett, 2006). Despite the popularity of the SJT, important questions still persist. One critical issue is the often-found difficulty of developing a SJT that measures one specific construct. This is reflected in the research literature, as substantial debate exists concerning what SJTs really measure. Does a SJT measure job knowledge (Schmidt & Hunter, 1993) or can a SJT be developed in such a way that it measures a specific construct? Several researchers (e.g., McDaniel, Morgeson, Bruhn Finnegan, Campion, & Braverman, 2001; Weekley & Jones, 1999) think the latter is possible, but only to a certain extent. Empirical evidence, namely, indicates that the constructs typically measured by SJTs are cognitive ability or *g*, conscientiousness, agreeableness, and emotional stability (McDaniel & Nguyen, 2001). In Chapter 6, it is questioned whether SJTs indeed can only measure *g*, conscientiousness, agreeableness, or emotional stability. To answer this question, a study is embarked upon to develop a video-based SJT measuring

Integrity, which is regarded as a central characteristic of police work. The construct validity of the SJT is examined both in the ethnic majority group and in the ethnic minority group. Furthermore, the score difference between the ethnic majority group and the ethnic minority group on this new type of test is investigated. The sixth and final research question consequently relates to the SJT and has been formulated as follows.

> *Research question 6:* To what extent does the SJT measure the same construct for the ethnic majority group compared to the ethnic minority group?

The six research questions as described above have guided the research that is presented in the five empirical chapters. Each chapter describes a separate study, which can be read independently from the other chapters. As a consequence, some overlap may exist across the Chapters 2 to 6 in the theory and method sections. Finally, in Chapter 7, answers to the research questions will be discussed. In closing, this chapter will present practical implications, recommendations, and ideas for future research.

Chapter 2

Applicant and method factors related to ethnic score differences in personnel selection: A study at the Dutch police [1]

The aim of this study was to examine applicant and method factors related to ethnic score differences on a cognitive ability test, a personality questionnaire, an assessment center (AC), an employment interview, and a final employment recommendation in the context of police officer selection (N = 13,526). Score differences between the majority group and the first-generation minority groups were comparable to research findings from the literature. However, score differences between the majority group and second-generation minority groups were much smaller. On the cognitive ability test and the personality questionnaire, most variability was explained by Dutch language-proficiency. Confirming assumed-characteristics theory, more variability on the interview and the employment recommendation was explained by Dutch language-proficiency and education than on the AC. Unsupportive of complexity-extremity theory, there seemed to be a general tendency to give lower scores to the ethnic minority group.

[1] This chapter was published as:
De Meijer, L. A. L., Born, M. Ph., Terlouw, G., & Van der Molen, H. T. (2006). Applicant and method factors related to ethnic score differences in personnel selection: A study at the Dutch police. *Human Performance, 19*(3), 219-251.
The study in this chapter was also presented at the 20th annual conference of the Society for Industrial and Organizational Psychology (SIOP), Los Angeles (CA), April 2005.

2.1 Introduction

The personnel selection literature has extensively investigated differences on psychological measures between ethnic minority and majority groups. This study focuses on ethnicity-related applicant demographics, such as language-proficiency and education, and their interplay with selection-method factors in their impact on test scores. Personnel selection measures can differ in the extent of assessor influence in the selection process. Those selection measures where there is no assessor influence (such as cognitive ability tests and personality questionnaires) are labeled as *objective measures*. By contrast, measures characterized by the involvement of an assessor (such as the assessment center and the employment interview) are labeled *subjective measures* (Bass & Barrett, 1981).

This study focuses on objective as well as subjective selection measures, with special attention paid to the subjective measures. First, an overview will be given of the literature findings stemming from North America and Europe on score differences between ethnic groups on objective measures, which will then be followed by a discussion on subjective measures and also the final employment recommendation to hire or not. Second, two theoretical perspectives developed within social psychology will be described. These perspectives concern the impact of perceptions of groups on evaluators' ratings, defined according to the ethnicity-related demographic characteristics of these groups. These are assumed-characteristics theory (Locksley, Borgida, Brekke, & Hepburn, 1980; Locksley, Hepburn, & Ortiz, 1982a; 1982b) and complexity-extremity theory (Linville, 1982; Linville & Jones, 1980). Hypotheses will then be derived and tested in the context of the Dutch police officer selection procedure over the last couple of years.

Objective Measures

Cognitive Ability Test

General cognitive ability, or *g*, has been found to be a consistent predictor of job performance across a variety of occupations (Schmidt & Hunter, 1998, 2004). This is especially the case for more complex job levels (Hunter, 1986). At the same time, several researchers (e.g., Goldstein, Zedeck, & Goldstein, 2002; Murphy, 2002; Outtz, 2002) have shown that the cognitive ability test represents the predictor most likely to have substantial adverse impact on employment opportunities for most ethnic minority groups. Ethnic score differences between .50 *SD* and 1.50 *SD* on cognitive ability tests have often been found (e.g., Herrnstein & Murray, 1994). However, evidence has been found that ethnic differences in cognitive ability test scores are considerably

larger than ethnic differences in measures of job performance (Hattrup, Rock, & Scalia, 1997; Waldman & Avolio, 1991).

Striving for a fully ethnicity-proof cognitive ability test has turned out to be a utopia. In such endeavors, a distinction has been made between tests that are more and tests that are less influenced by cultural aspects. Cattell (1987) made a well-known distinction between 'fluid' intelligence and 'crystallized' intelligence. Fluid intelligence relates to basic reasoning, which is necessary for problem solving, is dependent on neuronal efficiency, and is very hereditary (Bors & Forrin, 1995; Horn & Noll, 1997; Jensen, 1993; Plomin, 1988). Crystallized intelligence can be seen as a result of the action of fluid intelligence on a certain (cultural) environment, which is dependent on one's learning experience and on the perceived importance of certain abilities. Thus, crystallized intelligence can have different forms in different cultures. Most cognitive ability tests appeal to basic cultural knowledge and abilities (e.g., instructions and items written are in a certain language or the tests appeal to scholastic abilities [Van den Berg & Van Leest, 1999]). For Western ethnic majority group members with a comparable cultural and scholastic background, the appeal to basic cultural knowledge and abilities is not a problem. The required knowledge and abilities are 'overlearned' and, therefore, these group members have the basic knowledge and ability that is required in almost every situation. However, for ethnic minority members, the appeal to scholastic abilities for cognitive tests may indeed be a problem. Research in The Netherlands by Bleichrodt and Van den Berg (1995) has shown that it is not so much the period of residence in The Netherlands which impacts upon cognitive ability test scores in general and crystallized intelligence in specific, but much more the age of immigration.

Personality Questionnaire

The use of personality questionnaires to assess ethnic minority group members has been criticized as well. Although personality questionnaires are generally of adequate reliability and validity in different ethnic groups (Anderson & Ones, 2003), critics assume they are of limited use for assessing individuals in a certain country or area who have a limited knowledge of the spoken language and culture of that area (Te Nijenhuis, 1997). Yet, less research than on cognitive ability tests has been done to answer the question whether different ethnic groups exhibit different scores on personality questionnaires, and what has been done has found mixed results. Hough (1998) in the United States (U.S.), and Ones and Anderson (2002) in the United Kingdom (U.K.) reported ethnic group differences in the negligible to moderate range. In The Netherlands, ethnic score differences on personality questionnaires have been investigated by Van Leest (1997) and Te Nijenhuis,

Van der Flier, and Van Leeuwen (1997). Van Leest (1997) investigated ethnic score differences between majority group members and Turks. He found a mean difference of -.29 *SD*, ranging between -1.34 *SD* for Routine in Methods (i.e., avoidance of uncertainty) and 1.04 *SD* for Assertiveness, where positive values indicated the ethnic majority group scoring higher. Te Nijenhuis et al. (1997) found significantly higher mean scores for ethnic minorities on Neuroticism, varying between -.79 *SD* and -.43 *SD*, for Neurosomatism, ranging between -1.19 *SD* and -.28 *SD*, and for Social Conformity, ranging from -.78 *SD* to -.52 *SD*, where negative *d* values indicated ethnic minorities scoring higher. Lower mean scores for ethnic minorities were found for Extraversion, differing between .05 *SD* and .50 *SD* (where positive values indicated the ethnic majority group scoring higher). The differences between the ethnic majority and minority groups were larger for Turks and Moroccans than for Surinamese and Dutch Antillean groups. As regular jobs in general require emotional stability, sociability, and flexibility, utilizing personality questionnaire findings would imply lower hiring chances for many positions for ethnic minority groups (Te Nijenhuis et al., 1997).

Differences between ethnic groups on personality questionnaire scores seem to be much smaller than differences on cognitive ability tests. However, the reported differences on personality questionnaires seem unsystematic and therefore difficult to interpret. Reported ethnic group differences to the disadvantage of ethnic minorities are larger in The Netherlands than in the U.S. and U.K.. Relatively little has been published which provides explanations for such differences in findings.

When employers want to maximize the skill level of their employees on the one hand and diversify their workforce on the other hand, they are saddled with a dilemma. Both goals cannot be achieved simultaneously, because of existing subgroup differences in the results on objective measures. One solution has been sought in the use of face-valid simulations as selection tools in order to evaluate job-relevant knowledge, skills, and abilities that are both cognitive and non-cognitive (Schmitt & Mills, 2001). We now turn to several of these tools, which contain a subjective evaluative element by an assessor.

Subjective Measures

Assessment Center (AC)

ACs are mostly used for the selection of higher-level managerial jobs (Cascio, 1991; Heneman & Heneman, 1994). Past research has indicated that score differences between ethnic groups on the AC vary between .02 *SD* and .58 *SD* to the advantage of the ethnic majority group (e.g., Goldstein, Yusko,

Braverman, Smith, & Chung, 1998; Goldstein, Yusko, & Nicolopoulos, 2001). Findings until now have been mostly restricted to highly complex jobs and to specific North American ethnic groups (Cascio, 1991; Heneman & Heneman, 1994).

Employment Interview

The employment interview is probably the most commonly used selection tool (Huffcutt & Roth, 1998). Likewise, there has been a substantial amount of research examining ethnic score differences in the employment interview. Findings until now indicate that score differences between Blacks and Whites vary between .14 *SD* and .56 *SD* in favor of the ethnic majority group (Huffcutt & Roth, 1998; Motowidlo, Carter, Dunnette, Tippins, Werner, Burnett, et al., 1992; Roth, Van Iddekinge, Huffcutt, Eidson, & Bobko, 2002). In The Netherlands, Van den Berg (2001) found a difference of .23 *SD* between ethnic minorities and the ethnic majority, in favor of the majority group. Van den Berg reported that an important part of the variability in evaluations could be explained by language-proficiency of the applicant.

Final Employment Recommendation

Predictor information of several different selection measures needs to be combined to form a final employment recommendation of an applicant. Predictor information can be combined either mechanically (mechanical prediction) or judgmentally (clinical prediction). *Clinical prediction* refers to a procedure in which a judge puts data together using informal, subjective methods (Grove, Zald, Lebow, Snitz, & Nelson, 2000). *Mechanical prediction* refers to statistical prediction without the interference of a human evaluator (Grove et al., 2000). Although most final selection decisions are achieved through clinical prediction, it is found in numerous studies (for a review, see Grove et al., 2000) that mechanical prediction is either equal or superior to clinical judgment. Why does a mechanical combination of data yield better results than a judgmental combination? One possible explanation is that decision makers are more likely to add considerable error if they are allowed to judgmentally combine both subjective data (e.g., ACs or interviews) along with objective data (e.g., scores on the cognitive ability test or personality questionnaire). Their perceptions of an applicant may influence their evaluations and ultimately their decisions to select or reject an applicant (Bass & Barrett, 1981). Because of the existing ethnic score differences on both objective and subjective measures, the combination of these measures into a final employment recommendation is likely to yield ethnic differences as well. However, to our knowledge no research specifically directed to this issue has been done until now.

In sum, a review of the literature indicates that ethnic score differences are found on all selection measures discussed and are mostly in favor of the ethnic majority group. Differences on the cognitive ability test seem to be the largest and most consistent throughout all studies (between .50 *SD* and 1.50 *SD*). Differences that were found on the AC (between .02 *SD* and .58 *SD*) and on the employment interview (between .14 *SD* and .56 *SD*) are in favor of the ethnic majority group as well, although these are smaller than on the cognitive ability test. Research on personality questionnaires has also found differences between ethnic groups, but the results are mixed. Little attention has been given to why differences exist between ethnic groups on non-cognitive measures as well as why differences exist between selection tools. Finally, to our knowledge, no research has been done on final employment recommendations in which scores on various measures are combined.

In this chapter, ethnic score differences on the cognitive ability test, the personality questionnaire, the AC, the employment interview, and the employment recommendation are investigated. Furthermore, possible explanations for score differences between ethnic groups are searched. Before deriving hypotheses about ethnic score differences on objective and subjective measures, we first want to focus on two theoretical perspectives from social psychology pertaining to the subjective measures.

Theoretical Perspectives from Social Psychology

Two theoretical perspectives that address the influence of demographic information of individuals on evaluations by others are so-called assumed-characteristics theory (Locksley et al., 1980; Locksley et al., 1982a; 1982b) and complexity-extremity theory (Linville, 1982; Linville & Jones, 1980). Coleman, Jussim, and Kelley (1995), Jussim, Coleman, and Lerch (1987), and Jussim, Flemming, Coleman, and Kohberger (1996) have investigated assumed-characteristics theory and complexity-extremity theory in laboratory experiments. The present study investigates the applicability these two theories in a field setting. Both theories propose different processes to explain how ethnicity-related demographic information about individuals may influence evaluations by others. Assumed-characteristics theory, complexity-extremity theory, and their possible effects on applicant evaluations in the AC, the employment interview, and the employment recommendation will be discussed below.

Assumed-characteristics theory suggests that, based on knowledge about certain demographics of a group, people make assumptions about other characteristics of this group. For example, on the basis of knowledge of a group's ethnicity, people assume that this group will have a certain socio-

economic status (SES), education, or personality. Assumed-characteristics theory suggests that members of an in-group will believe their own characteristics are more favorable than characteristics of members of an out-group (Coleman et al., 1995; Jussim et al., 1987; 1996). Having more relevant ethnicity-related demographic information about out-group members should nevertheless substantially decrease the unfavorable assumed characteristics and evaluations of out-group members should become more positive. In other words, this theory supposes that the new ethnicity-related demographic information of the applicant is more positive than the assumed characteristics on the basis of ethnicity. When in-group members have relevant ethnicity-related demographic information about out-group members, e.g., information about someone's education and language-proficiency, this information should diminish the negative group membership effects on the basis of, e.g., ethnicity (Jussim, 1990, 1991, 1993; Locksley et al., 1980; Rokeach & Mezel, 1966; for a review see Swim, Borgida, Maruyama, & Myers, 1989). Assumed-characteristics theory would lead to the following expectations: Even if some factors, such as the knowledge of someone's ethnicity, lead to an evaluation in favor of individuals belonging to one group, other relevant ethnicity-related demographic information – information other than someone's ethnicity, i.e., someone's education or language-proficiency – of that person should be more influential than group membership on the basis of ethnicity (Coleman et al., 1995; Jussim et al., 1987; 1996). Furthermore, if ethnicity-related demographic information such as education and language-proficiency is not available, factors such as ethnicity and the negative out-group membership effects of them will have a stronger influence on evaluations of individual assessors.

Complexity-extremity theory (Linville, 1982; Linville & Jones, 1980) starts with the assumption that people have more contact with in-group members than with out-group members. Because in-group members have more contact with other in-group members, they will develop more complex representations of in-group members than of out-group members. When observers use these more complex representations to evaluate an in-group member, they are likely to give accurate evaluations. When complex representations are developed, an observer has knowledge about both good and bad characteristics about the in-group member, which renders an extreme evaluation unlikely. Low complex or simple representations will be developed about out-group members. When these simple representations are used in evaluating an out-group member, extreme evaluations are more likely because the out-group member can more easily be seen as all good or all bad. Thus, a complex representation of someone will lead to less chance of extremity in evaluations, and a simple representation leads to a higher chance of extremity in evaluations (Coleman et al., 1995; Jussim et al., 1987; 1996). Complexity-

extremity theory would lead to the following judgmental outcomes: (a) an out-group member whose demographics can be seen as positive (e.g., high education) will be evaluated extremely favorable – even more favorable than an in-group member with those same demographics (Coleman et al., 1995; Jussim et al., 1987; 1996); and (b) an out-group member whose demographics can be seen as negative (e.g., low education) will be evaluated extremely unfavorable – even more unfavorable than an in-group member with those same demographics (Coleman et al., 1995; Jussim et al., 1987; 1996).
We will employ assumed-characteristics theory and complexity-extremity theory as potential explanatory frameworks for assessors' evaluations on the subjective measures, i.e., the dimensions that are assessed in the AC, the employment interview, and the employment recommendation.

Having now discussed the literature available on ethnic score differences on various selection tools and some possible explanations for these differences, hypotheses are formulated in the following section.

Overview of Hypotheses
The hypotheses may be divided in three groups. First, consistent with findings from the literature, it is expected that differences between ethnic minorities and ethnic majorities to the advantage of the majority group exist on all selection tools, i.e., the cognitive ability test, the personality questionnaire, the AC, the employment interview, and the employment recommendation (Hypothesis 1a). Further, based on the literature it is expected that ethnic score differences will be largest on the cognitive ability test and lowest on the personality questionnaire (Hypothesis 1b). Research by Bleichrodt and Van den Berg (1995) has shown that first-generation ethnic minority group members who moved to The Netherlands from countries such as Turkey, Morocco, Surinam, or the Dutch Antilles before the age of seven (before starting their primary education), score significantly higher on cognitive ability tests than first-generation ethnic minorities who moved to The Netherlands after the age of seven. The scores of first-generation ethic minority group members who moved to The Netherlands before the age of seven were still lower, though, than scores of ethnic majority group members (Bleichrodt & Van den Berg, 1995). Second-generation minority group members, in contrast to first-generation minorities, are born in The Netherlands and, therefore, will have passed through the Dutch educational system. First-generation ethnic minority members are born outside The Netherlands. Large numbers of first-generation minorities did not receive their education in The Netherlands. Second-generation ethnic minorities do not only differ from first-generation ethnic minorities in terms of education. Second-generation ethnic minorities, because they are born in The

Netherlands, are also confronted with the Dutch culture and its norm and values to a larger extent than first-generation ethnic minorities. However, most of the second-generation ethnic minority families still speak their native language at home and, to some extent, have their own customs. Therefore, second-generation minorities are not yet fully integrated into Dutch society and they still differ from the Dutch majority (Weijters & Scheepers, 2003). Extending the findings of Bleichrodt and Van den Berg (1995), it is expected that score differences between first- and second-generation ethnic minority groups to the advantage of the second-generation minority group exist on all selection tools (Hypothesis 1c). In line with findings from Bleichrodt and Van den Berg (1995) on differences in crystallized intelligence and fluid intelligence, it is expected that on subtests measuring crystallized intelligence (Cattell, 1987) the differences between first- and second-generation minority groups will be larger than on subtests measuring fluid intelligence (Hypothesis 1d).

A second and a third group of hypotheses will be addressed below. A lot of research has been done to explain differences on cognitive ability tests between ethnic groups in North America. There has often been a tendency in existing research, to treat ethnic minorities as a homogeneous group that merely contrasts with the ethnic majority group. That is, a dichotomous distinction is made between Whites and non-Whites or between the majority and the minority group. This approach ignores the many visible and cultural differences between ethnic groups that may affect scores on selection instruments. The main ethnic minority groups in North America are Blacks, Hispanics/Latinos and Asians. These American ethnic minority groups moved to North America generations ago, whereas in Europe ethnic minority groups mainly moved to European countries from the 1960s onward. Therefore, first- and second-generation ethnic minority groups are at the center of attention in European research on ethnic group differences. Because of the difference between the length of residence of ethnic minorities in North-America and in Europe, the language-proficiency of ethnic minority samples in North American research is probably better than the language-proficiency of ethnic minority samples in European research. Explanations for ethnic cognitive ability differences in North America are often searched in the context of SES and background characteristics, whereas in The Netherlands, where most research focuses on Antillean, Moroccan, Surinamese, and Turkish minority groups, group differences are sought in Dutch language-proficiency and being a first- or second-generation minority (Bleichrodt & Van den Berg, 1995).

The second group of hypotheses is aimed at investigating to what extent

ethnic score differences on the objective and subjective measures can be explained by the following applicant demographics: (1) Dutch language-proficiency; (2) education; and (3) ethnicity. It is hypothesized that as the objective tests, i.e., the cognitive ability test and the personality questionnaire are tests written in Dutch, Dutch language-proficiency will explain more of the variability between ethnic groups than education and ethnicity (Hypothesis 2a).

Hypotheses 2b and 2c are derived from assumed-characteristics theory (Coleman et al., 1995; Jussim et al., 1987, 1996). The hypotheses are aimed at investigating to what extent ethnic score differences on the employment interview and the employment recommendation on the one hand and the AC on the other hand can be explained by the following applicant ethnicity-related demographics: (1) Dutch language-proficiency; (2) education; and (3) ethnicity. In the employment interview and the final recommendation, the assessor has knowledge of the applicant's language-proficiency, education, and ethnicity. In the AC, no such knowledge is given to the assessors. The reason why assessors do have knowledge about demographic information of the applicant during the interview and the employment recommendation and assessors do not have this knowledge during the AC is that interviewers also write the final recommendation and all information about a certain applicant is at the interviewers' disposal.

From assumed-characteristics theory it is hypothesized that for the employment interview and the final recommendation, applicant ethnicity-related demographics, namely Dutch language-proficiency and education, will explain more of the variability in assessors' evaluations than ethnicity itself (Hypothesis 2b). Furthermore, it is hypothesized that for the AC, Dutch language-proficiency and education will not explain more of the variability in assessors' evaluations but as much as or less than ethnicity (Hypothesis 2c).

The third group of hypotheses is derived from complexity-extremity theory (Coleman et al., 1995; Jussim et al., 1987, 1996). It is hypothesized that ethnic majority assessors will evaluate ethnic minority applicants with an excellent Dutch language-proficiency and education higher on the interview and the employment recommendation than ethnic majority applicants with the same excellent Dutch-language-proficiency and education (Hypothesis 3a); and that ethnic majority assessors will evaluate ethnic minority applicants with a low Dutch language-proficiency and education lower on the interview and the employment recommendation than ethnic majority applicants with the same low Dutch-language-proficiency and education (Hypothesis 3b).

2.2 Method

Participants and Procedure

Data came from a first-generation minority group, a second-generation minority group, and a majority group, who applied for a position at the Police Academy of The Netherlands from September 2001 until July 2003. The largest first- and second-generation ethnic minority groups are from the Dutch Antilles, from Morocco, Surinam, and Turkey. The dataset consisted of 11,432 applicants. Data of 11,409 applicants were used, of which 672 applicants were first-generation ethnic minorities and 734 applicants were second-generation ethnic minorities. Data of 23 applicants were incomplete. These cases were removed from the dataset. The professions for which accepted students were to be trained for were assistant police employee, police employee, or all-round police employee. Applicants who were interested in a job as police officer first applied to the local police force where they wanted to work after they would complete their training. For the selection procedure, the local police forces routinely send all applicants to the National Police Center for Competence Assessment and Monitoring (CCM). During a requirement check at the CCM, the following minimal criteria are checked on the basis of an application form: minimal age (16 years), Dutch nationality (first or second), possession of a swimming diploma, no criminal record, possession of a school diploma (minimal level is preparatory vocational education level B [VBO-B]). Applicants in the selection process went through two stages. During the first stage a Dutch language-proficiency test was filled out. During the second phase a physical exercise, a cognitive ability test, a personality questionnaire, an AC assignment and an employment interview were executed. The psychologist who conducts the interview is also the one who writes the final employment recommendation to the police force. For the employment recommendation, the test results of the personality questionnaire, the AC ratings, and the employment interview ratings are used. Next to the final recommendation, the final dossier to the local police forces exists of test scores of the physical exercise, the cognitive ability test, and the language-proficiency test (for an overview of the selection procedure: see Appendix A).

Table 1 shows the distributions of the groups in terms of demographic variables. The ethnic minority group from countries classified as other (N = 325, 2.8% of total group) consists mostly of people from Eastern and Western Europe and Western non-European (e.g., U.S., Canada, and Australia) countries (72% of others), but also from Asia (with the exception of Turkey, Japan, and the Dutch Indies), Africa (with the exception of

Morocco), South America (with the exception of Surinam), and Central America (with the exception of the Dutch Antilles). In view of the heterogeneity of this group, its data are only used to test hypotheses 2 (a-c) and 3 (a-b). To test these hypotheses the ethnic minority groups are taken together.

Within all ethnic groups the largest number of applicants were male, especially within the first-generation minority group (mean % male = 75). Within the majority group 66% were male. The mean age of the applicants of the first-generation minority group (M = 28.00; SD = 7.05) was higher than the mean age of the second-generation minority group (M = 21.85; SD = 4.57; t = 19.21, $p < .05$) and of the majority group (M = 23.92; SD = 7.11; t = 14.49, $p < .05$). The largest percentage of applicants was within the majority group (88%). Six percent of the applicants were first-generation minority members and 6% were second-generation applicants. For 24 applicants it was not known if they were first-generation ethnic minority, second-generation-ethnic minority, or majority group members.

All assessors (82 conducting the interview and the employment recommendation and 116 conducting the AC) in the selection process of the CCM had a background as vocational advisor or psychologist. Eighty-six percent of the assessors in the interview and the final recommendation were female and 78% of the assessors in the AC were female. Nearly all assessors were majority group members and all had a high educational level (higher professional education [HBO], or research-oriented education [WO]).

Table 1

Distribution of Majority Group Members and First- and Second-Generation Minority Group Members in Terms of Gender, and Age

Position	*n* (%)	% male	mean age (*SD*)
1st Generation minority group	672 (5.9)	75	28.00 (7.05)
2nd Generation minority group	734 (6.4)	72	21.85 (4.57)
Majority group	10,003 (87.5)	66	23.92 (7.11)
Total	11,432 (100)	67	24.03 (7.06)

	1st generation minority group			2nd generation minority group		
Ethnicity	*n* (%)	% male	mean age (*SD*)	*n* (%)	% male	mean age (*SD*)
Dutch Antilles	75 (0.7)	75	29.58 (6.97)	33 (0.3)	73	21.78 (4.72)
Morocco	90 (0.8)	74	27.57 (6.74)	110 (1.0)	71	21.16 (3.28)
Surinam	129 (1.1)	81	28.93 (6.86)	123 (1.1)	69	21.95 (5.08)
Turkey	162 (1.4)	74	27.54 (6.85)	334 (2.9)	75	21.23 (3.25)
Other	216 (1.9)	72	27.41 (7.38)	109 (1.0)	66	23.63 (6.31)

Note. Of 24 (0.2%) applicants it was not known if they were first-generation ethnic minority, second-generation ethnic minority, or majority group members and the ethnicity of 25 applicants (0.2%) could not be determined.

Measures

Cognitive Ability

The Police Intelligence Test (PIT; Rijks Psychologische Dienst, 1975) is a cognitive ability test and consists of 107 items divided over 6 subtests: Analogies (verbal comprehension), Arranging Pictures (picture arrangement), Series of Numbers (numerical reasoning), Silent Reading (word fluency), Folding Figures (spatial ability), and Series of Figures (inductive reasoning). The time limit is 51 minutes. Applicants completed the PIT in Dutch. Prior research by Lem and Van Doorn (2000) indicated alpha reliabilities varying from .69 for Series of Numbers, to .87 for Folding Figures. The correlations between the subscales varied from .32 to .57. A study by Van der Maesen (1992) showed corrected predictive validity coefficients of .39 and .46 (N =162).

Personality

To measure the Big Five factors Extraversion, Altruism, Conscientiousness, Emotional Stability, and Intellect, the Police Personality Questionnaire (PPV; Van Leeuwen, 2000) was used. The applicants completed the PPV in Dutch. A recent progress report by Klinkenberg and Van Leeuwen (2003) indicated alpha reliabilities varying from .72 for Conscientiousness, to .78 for Intellect. Correlations between the scales are all lower than .60. Comparison with NEO-PI-R showed observed construct validity coefficients between .17 and .58 (N = 160). A study by Lem and Van Doorn (2000) showed observed predictive validity coefficients between .15 and .43 (N = 61).

Assessment Center (AC)

A role-play exercise is utilized, in which an assessor and an actor independently make ratings on a 7-point Likert-scale ranging from 1 (extremely weak) to 7 (excellent), on each of the following seven dimensions: Communication Skills, Social Skills, Empathy, Initiative, Stress Tolerance, Authority, and Decisiveness. Interrater reliabilities ranged from .82 to .88 (N = 198). Principal component analysis with varimax rotation yielded two factors, Agency and Communion (in accordance with Wiggins and Trapnell, 1996), which together explained 77% of the variance. As a measure of Agency, the average rating across the dimensions of Authority, Decisiveness, Initiative, Communication Skills, and Stress Tolerance was used ($\bar{r}$ = .59; α = .87). As a measure of Communion, the average rating of the dimensions Social Skills and Empathy was used ($\bar{r}$ = .77; α = .87). The reliability of the difference (r_{diff}) between scores on Agency and Communion was .78.

Employment Interview

The interview questions are focused on evaluating behavior on the following eight dimensions: Communication Skills, Social Skills, Flexibility, Stress Tolerance, Emotional Stability, Tolerance Towards Others, Integrity, and Self-Understanding. A single interviewer conducts the interview. The interviews are semi-structured and behaviorally based, with one behaviorally anchored 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent) for each of the eight dimensions. The average rating across the eight dimensions was used as the dependent variable because the ratings were substantially correlated ($\bar{r}$ = .42; α = .85). Moreover, principal component analysis with varimax rotation yielded one interview factor that explained 50% of the variance.

Final Employment Recommendation

The final recommendation as to whether an applicant is fit for a job as police officer is based on results from the personality questionnaire (PPV), the AC, and the employment interview. These scores are integrated into an employment recommendation. The dimensions in the final recommendation are: Communication Skills, Social Skills, Empathy, Initiative, Flexibility, Stress Tolerance, Authority, Decisiveness, Tolerance Towards Others, Integrity, and Self-Understanding (for definitions, see Appendix B). A 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent) is used to evaluate the behavior on the eleven dimensions. Principal component analysis with varimax rotation yielded three employment-recommendation factors, Agency, Communion, and Socio-Cultural Awareness, which altogether explained 67% of the variance. As a measure of Agency, the average rating across the dimensions Authority, Decisiveness, Initiative, Communication Skills, Stress Tolerance, and Flexibility was used ($\bar{r}$ = .48; α = .85). As a measure of Communion, the dimensions Social Skills and Empathy, were used ($\bar{r}$ = .66; α = .79) and for Socio-Cultural Awareness the dimensions ($\bar{r}$ = .39; α = .65), Tolerance Towards Others, Integrity, and Self-Understanding. The reliability of the difference (r_{diff}) between scores on Agency and Communion is .51, r_{diff} between scores on Agency and Socio-Cultural Awareness is .58, and r_{diff} between scores on Communion and Socio-Cultural Awareness is .57.

Analyses

First Group of Hypotheses

Results from preliminary analyses showed that all measures were found to be structural equivalent (for detailed information, please contact the author). Levene's tests for equality of variances and *t* tests for equality of means were conducted to index ethnic group differences on the various selection

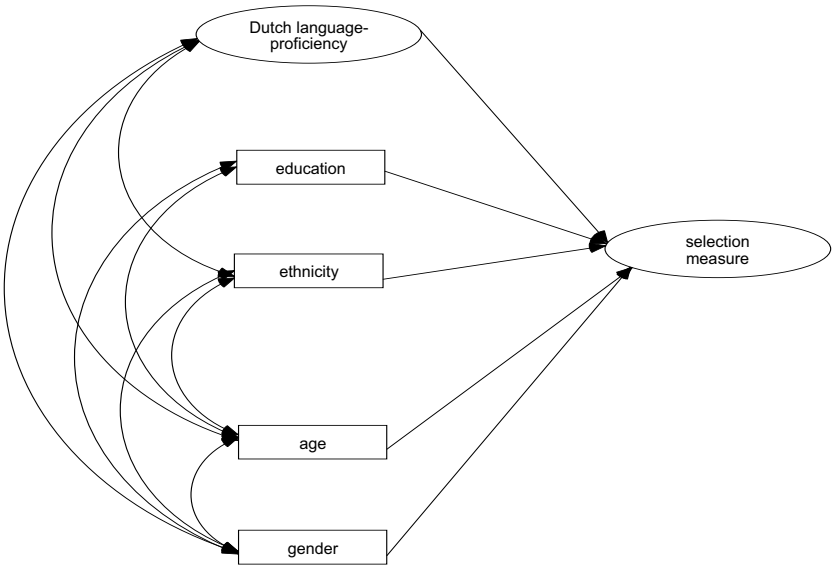
measures (Hypothesis 1a). Following Ones and Anderson (2002), standardized effect sizes (*d* values) between the means of the various groups of interest were computed to get an indication of the magnitude of the group differences on the various selection instruments irrespective of sample size (Hypotheses 1 b-d). *D* values index the standardized mean differences between any two groups being compared (Cohen, 1988). Positive *d* values indicate higher mean scores for the majority group and negative *d* values indicate higher mean scores of a minority group (Antillean, Moroccan, Surinamese, or Turkish group). Although effect sizes can theoretically range between positive and negative infinity, Cohen (1988) suggests that effect sizes of about .20 in magnitude are small, around .50 are medium, and above .80 are large. To conduct Levene's tests and *t* tests and to compute *d* values, observed differences on dimensions scores were used that were uncorrected for age, gender, and education. Corrected *d* values only differed marginally (about .01 *SD*) from uncorrected *d* values.

Second Group of Hypotheses

Structural equation modeling (SEM) with Amos 5.0 (Arbuckle, 2003) was used to investigate to what extent score differences on the cognitive ability test, the personality questionnaire (the objective measures), the AC, the employment interview, and the final recommendation (the subjective measures) between ethnic groups could be explained by a number of factors (Hypotheses 2 a-c). These factors are: (1) Dutch language-proficiency, (2) education, and (3) ethnicity. All factors are ordinal variables. An ordinal conception of ethnicity manifests itself in a hierarchy of ethnic groups in terms of social distance from the Dutch majority (Hraba, Hagendoorn, & Hagendoorn, 1989). Several studies have found consensus on the hierarchy of ethnic groups in The Netherlands (e.g., Hraba et al., 1989; Verkuyten, Hagendoorn, & Masson, 1996) where European groups were placed on top, followed by colonial and then Islamic groups at the bottom. More specifically, the following hierarchy is used (Hraba et al., 1989): (1) Dutch majority, (2) Western ethnic minority (which includes people from Western and Eastern Europe, and Western non-European countries), (3) Dutch Antilles, (4) Surinam, (5) Morocco, and (6) Turkey.

Because the factors Dutch language-proficiency and ethnicity had a moderate intercorrelation (r = .37; education and ethnicity, and language-proficiency and education did not correlate), a general model was created which took the intercorrelation between Dutch language-proficiency and education into account (see Figure 1). For measuring Dutch language-proficiency, a Dutch language test (IBO; Bureau Interculturele Evaluatie, 2000) was used that had previously turned out to be very useful in the practice of educational

institutes.



Note. Selection measures are the cognitive ability test, the personality questionnaire, the AC, the employment interview, and the final employment recommendation. Age and gender are control variables.

Figure 1. Path model to test the explanatory power of Dutch language-proficiency, education, and ethnicity

With regard to Hypothesis 2a, the effects of Dutch language-proficiency, education, and ethnicity were examined on cognitive ability. In order to examine this, a specific model was created where *g*-loaded subtests – subtests that measure fluid intelligence – were used as control variables. This was done

because Dutch language proficiency was intercorrelated with the *g*-loaded cognitive ability subtests ($r = .67$).

Third Group of Hypotheses

To test complexity-extremity theory (Hypotheses 3 a-b), Levene's tests for equality of variances and *t* tests for equality of means were conducted. Also, standardized effect sizes (*d* values) were calculated to get an indication of the magnitude of the group differences. Positive *d* values indicate higher mean scores for the majority group. Because of the small first- and second-generation sample sizes operationalized as 'high' and 'low', comparisons were only made between the ethnic majority group and the undifferentiated ethnic minority group of applicants with a high Dutch-language-proficiency and education and applicants with a low Dutch language-proficiency and education. The AC was not used for testing complexity-extremity theory because in the AC, as said earlier, no information was given to the assessors on Dutch language-proficiency and education. Age, gender, and cognitive ability were used as control variables because the aim, here, was not directed at age, gender, and cognitive ability differences.

2.3 Results

First Group of Hypotheses

The results relevant to the Hypotheses 1 (a-d) are presented in Table 2. Consistent with the findings from the literature, significant score differences between the ethnic majority group and ethnic minority groups to the advantage of the majority group, existed on all selection tools (Hypothesis 1a). The only exception was the personality questionnaire (PPV) dimension Conscientiousness, where minority groups systematically scored higher than the majority group.

In accordance with Hypothesis 1b, score differences between the ethnic majority group and the ethnic minority groups were largest on the cognitive ability test (PIT) and lowest on the personality questionnaire (PPV). Score differences on the PIT varied from *d* values of .06 *SD* ($t = .38$, *ns*) on Spatial Ability (PIT), to 1.30 *SD* ($t = 12.30$, $p < .001$) on Inductive Reasoning (PIT). Score differences on the PPV ranged between -.49 *SD* ($t = -4.94$, $p < .001$) on Conscientiousness, and .65 *SD* ($t = 5.62$, $p < .001$) on Extraversion.

Table 2

Differences between the Majority Group and First- and Second-Generation Minority Groups on a Cognitive Ability Test (PIT), a Personality Questionnaire (PPV), an AC, an Employment Interview, and a Final Employment Recommendation

Dimension	*d*				*t* test (*t*) Levene's test (*F*)			
	Majority vs. first-generation minority groups							
PIT	A	M	**S**	*T*	A	M	**S**	*T*
Verbal Comprehension	1.27	.98	**1.00**	*1.18*	11.06**	9.28**	**11.41****	*14.25***
					.08	1.60	**1.25**	*3.62*
Inductive Reasoning	.90	1.30	**1.00**	*.96*	6.56**	12.30**	**11.35****	*10.66***
					6.48*	.41	**1.51**	*8.07**
Numerical Reasoning	.66	1.08	**.79**	*.67*	5.67**	13.61**	**10.76****	*8.49***
					3.38	13.88**	**10.42***	*3.23*
Word Fluency	1.03	1.15	**.88**	*1.12*	11.17**	10.76**	**9.89****	*17.12***
					6.16*	1.72	**1.54**	*13.65***
Spatial Ability	.69	1.06	**.79**	*.87*	5.91**	9.94**	**8.95****	*10.11***
					1.01	.45	**2.33**	*3.90**
Picture Arrangement	.90	.98	**.98**	*1.11*	6.15**	7.45**	**8.79****	*12.31***
					14.68**	15.07**	**25.17****	*6.68**
PPV								
Extraversion	.65	.13	**.41**	*.19*	5.62**	1.22	**4.62****	*2.39**
					3.51	2.93	**1.59**	*.05*
Altruism	.25	.01	**.16**	*.00*	1.92*	.08	**1.65**	*-.05*
					6.82*	6.73*	**5.84***	*10.95**
Conscientiousness	-.49	-.30	**-.52**	*-.43*	-4.94**	-2.88*	**-4.72****	*-5.47***
					8.66*	.03	**2.09**	*2.67*
Emotional Stability	.04	.11	**-.07**	*-.28*	.39	1.02	**-.79**	*-3.49***
					.01	.07	**.14**	*1.18*

Intellect	.35	.21	**.30**	*.44*	3.04*	1.66	**3.13***	*5.54***
					.53	14.65**	**5.35***	*1.69*
AC								
Agency	.56	.39	**.43**	*.46*	4.27**	3.38**	**4.67****	*4.09***
					1.54	.00	**.16**	*1.56*
Communion	.27	.29	**.35**	*.30*	2.31*	2.49*	**3.73****	*3.34***
					5.44*	.03	**.57**	*2.89*
Interview	.50	.23	**.46**	*.34*	3.87**	2.05*	**4.99****	*3.53***
					.44	.33	**.48**	*3.96**
Final Recommendation								
Agency	.59	.37	**.55**	*.47*	4.47**	3.15*	**5.84****	*5.17***
					.34	.18	**1.00**	*.20*
Communion	.35	.27	**.39**	*.17*	2.63*	2.35*	**4.21****	*1.74**
					.06	1.20	**.28**	*4.08**
Socio-Cultural Awareness	.27	.21	**.33**	*.19*	2.01*	1.77*	**3.35****	*2.06**
					2.53	1.20	**.81**	*3.58*
	Majority vs. second-generation minority groups							
PIT	A	M	**S**	*T*	A	M	**S**	*T*
Verbal Comprehension	.33	.48	**.41**	*.82*	1.89*	4.36**	**3.91****	*14.90***
					.00	6.28*	**6.53***	*.87*
Inductive Reasoning	.17	.62	**.51**	*.68*	.95	6.47**	**4.84****	*10.84***
					.17	2.36	**6.29***	*11.41**
Numerical Reasoning	.43	.56	**.58**	*.43*	2.46*	6.69**	**7.46****	*8.50***
					1.65	4.01*	**5.75***	*4.15**
Word Fluency	.28	.60	**.48**	*.71*	1.61	6.22**	**5.32****	*14.12***
					.13	2.42	**3.71**	*7.42**
Spatial Ability	.06	.63	**.33**	*.55*	.38	5.63**	**3.64****	*9.98***
					1.86	4.39*	**.09**	*.04*

Picture Arrangement	.16	.47	**.50**	*.74*	.93	4.87**	**4.79****	*12.29***
					.01	2.67	**5.72***	*10.77**
PPV								
Extraversion	-.03	-.05	**.03**	*.08*	-.16	-.51	**.34**	*1.47*
					5.88*	.00	**.21**	*.15*
Altruism	.14	-.04	**.03**	*-.06*	.79	-.40	**.29**	*-.96*
					.22	4.50*	**7.68***	*7.09**
Conscientiousness	-.11	-.23	**-.15**	*-.45*	-.62	-2.39*	**-1.55**	*-8.79***
					.11	.00	**3.95***	*12.34***
Emotional Stability	.05	-.17	**-.09**	*-.20*	.27	-1.80*	**-1.04**	*-3.78***
					2.77	.35	**.90**	*4.56**
Intellect	.06	.17	**.02**	*.38*	.34	1.59	**.23**	*6.86***
					.79	4.62*	**3.79**	*1.58*
AC								
Agency	.24	.32	**.21**	*.59*	1.36	3.27**	**2.16***	*10.29***
					.14	.72	**1.15**	*.00*
Communion	.19	.24	**.02**	*.36*	1.09	2.38*	**.22**	*6.28***
					1.01	.03	**.02**	*.66*
Interview	.24	.15	**.24**	*.62*	1.39	1.56	**2.58***	*10.72***
					2.08	1.19	**.23**	*.46*
Final Recommendation								
Agency	.32	.32	**.26**	*.68*	1.80*	3.25**	**2.78***	*11.94***
					.99	.19	**.92**	*.07*
Communion	.11	.18	**.09**	*.35*	.61	1.85*	**.94**	*6.15***
					.09	.05	**.30**	*2.50*
Socio-Cultural Awareness	.36	.09	**.29**	*.41*	1.99*	.72	**2.98***	*7.06***
					.66	4.22*	**.01**	*1.74*

	First-generation vs. second-generation minority groups							
PIT	A	M	**S**	*T*	A	M	**S**	*T*
Verbal Comprehension	.90	.44	**.53**	*.29*	4.81**	3.18*	**4.35****	*3.02**
					.02	.44	**.96**	*.92*
Inductive Reasoning	.64	.60	**.44**	*.25*	3.21*	4.45**	**3.61****	*2.64**
					3.28	.22	**.65**	*.13*
Numerical Reasoning	.26	.61	**.25**	*.26*	1.27	4.53**	**2.01***	*2.76**
					.03	2.61	**.31**	*.10*
Word Fluency	.81	.60	**.44**	*.46*	4.22**	4.36**	**3.56****	*4.89***
					1.29	.02	**.44**	*2.47*
Spatial Ability	.49	.37	**.42**	*.31*	2.42*	2.64*	**3.39****	*3.22***
					.30	.59	**.61**	*2.16*
Picture Arrangement	.61	.45	**.40**	*.34*	3.09*	3.24**	**3.29****	*3.54***
					3.46	3.35	**2.84**	*.06*
PPV								
Extraversion	.60	.15	**.36**	*.10*	2.71*	1.09	**2.79***	*1.06*
					10.10*	1.29	**.22**	*.01*
Altruism	.08	.03	**.11**	*.05*	.36	.25	**.88**	*.47*
					1.05	.19	**.10**	*1.35*
Conscientiousness	-.47	-.09	**-.27**	*.01*	-2.31*	-.64	**-2.17***	*.11*
					2.85	.03	**6.46***	*.39*
Emotional Stability	-.02	.27	**.02**	*-.08*	-.12	1.89*	**.13**	*-.83*
					1.96	.29	**.82**	*.10*
Intellect	.24	.03	**.25**	*.05*	1.18	.20	**2.01***	*.55*
					.13	1.86	**.04**	*2.75*
AC								
Agency	.34	.09	**.23**	*-.12*	1.33	.41	**1.70***	*-1.25*
					.18	.32	**.22**	*.99*
Communion	.13	.06	**.32**	*-.05*	.39	.31	**2.40***	*-.56*
					.32	.05	**.18**	*.94*
Interview	.29	.08	**.21**	*-.24*	1.17	.50	**1.59**	*-2.36**
					2.60	.07	**.02**	*1.53*

Final Recommendation								
Agency	.29	.06	**.28**	*-.21*	1.15	.26	**2.03***	*-2.01**
					.20	.00	**.00**	*.05*
Communion	.25	.09	**.30**	*-.17*	1.04	.60	**2.30***	*-1.58*
					.01	1.14	**.59**	*.67*
Socio-Cultural Awareness	-.08	.11	**.05**	*-.21*	-.48	.72	**.24**	*-1.95**
					.14	.22	**.29**	*3.95**

Note. A means the Antillean group, M means the Moroccan group, **S** means the Surinamese group, and *T* means the Turkish group. A positive *t* values means that the mean of the majority group is higher than the mean of the minority group.

d = difference between majority and minority ethnic group means in standard deviation units, positive *d* values indicate higher ethnic majority groups scoring higher.

* $p < .05$, ** $p < .001$ (two-tailed for Levene's test and one-tailed for the *t* test).

Hypothesis 1c predicted that differences between the first-generation minority groups and the second-generation minority groups to the advantage of the second-generation minority groups exists on all selection tools. To test this hypothesis, first- and second-generation ethnic minority groups were compared. Positive *d* values indicate the second-generation minority group scoring higher than the first-generation minority group. For the cognitive ability test (PIT), 96% of the comparisons supported the hypothesis. Less support was found on the personality questionnaire (PPV; 20%), the AC (25%), the employment interview (0%), and the employment recommendation (25%). Differences varied between -.47 *SD* (t = -2.31, p < .05) on Conscientiousness (PPV) for the difference between the first- and second-generation Antillean group, and .90 *SD* (t = 4.81, p < .001) on Verbal Comprehension (PIT), also for the difference in the Antillean group. Three remarkable findings are highlighted. Firstly, on the PPV dimension Conscientiousness, the difference between the first- and second-generation Antillean and Surinamese groups was to the advantage of the first-generation Antillean (-.47 *SD*; t = -2.31, p < .05) and Surinamese group (-.27 *SD*; t = -2.17, p < .05). Secondly, the Turkish group showed a different pattern. Scores on the interview (-.24 *SD*; t = -2.36, p < .05) and the employment recommendation (the dimension Agency [-.21 *SD*; t = -2.01, p < .05] and the dimension Socio-Cultural Awareness [-.21 *SD*; t = -1.95, p < .05]) showed differences to the advantage of the first-generation Turkish group. Lastly, as shown in Table 2 ('majority vs. second-generation minority'), the majority group still scored higher than the second-generation minority groups. The personality questionnaire scores (PPV), again, showed different results.

The results relevant to Hypothesis 1d showed score differences on all subtests of the cognitive ability test between the first-generation minority groups and the second-generation minority groups. All differences were to the advantage of the second-generation minority group. To further look at the results, a distinction was made between subtests for crystallized intelligence and subtests for fluid intelligence. Subtests that measure crystallized intelligence are Verbal Comprehension, Numerical Reasoning, Word Fluency, and Picture Arrangement. Subtests that measure fluid intelligence are Inductive Reasoning and Spatial Ability. In line with findings from Bleichrodt and Van den Berg (1995), the score differences between both generations on subtests of fluid intelligence were somewhat smaller than the differences on subtests of crystallized intelligence. The differences of fluid intelligence varied from .25 *SD* (t = 2.64, p < .05) to .64 *SD* (t = 3.21, p < .05; mean difference is .44 *SD*). For crystallized intelligence, the differences varied from .25 *SD* (t = 2.01, p < .05) to .90 *SD* (t = 4.81, p < .001; mean difference is .48 *SD*).

Second Group of Hypotheses

The results relevant to hypotheses 2 (a-c) are shown in Tables 3 and 4. From the fit indices shown in Table 3, it can be concluded that the model fit (χ^2/df of 55.88 and 54.52; TLI of .88; CFI of .92; RMSEA of .07) was good. Hypothesis 2a stated that Dutch language-proficiency could explain more of the variability in ethnic score differences on the cognitive ability test (PIT) and the personality questionnaire (PPV) than education and ethnicity. Table 4 reports the unstandardized and standardized path coefficients. Support for this hypothesis was found on the PIT as well as the PPV. On the cognitive ability test (PIT), the explained variance by Dutch language-proficiency was 16% (unstandardized path coefficient of .40, $p < .001$) compared to 0.05% by education (unstandardized path coefficient of .07, $p < .001$) and 0.05% by ethnicity (unstandardized path coefficient of .10, $p < .001$). For the personality questionnaire (PPV), support for hypothesis 2a was found but the support was less overwhelming than for the cognitive ability test. Dutch language-proficiency explained more variance than education and ethnicity, accounting for 0.60% (unstandardized path coefficient of .09, $p < .001$) of the variability in test scores, whereas education (unstandardized path coefficient of -.01, *ns*) and ethnicity (unstandardized path coefficient of -.01, *ns*), together, accounted for only 0.02% of the variability.

Hypothesis 2b, derived from assumed-characteristics theory, predicted that Dutch language-proficiency and education together would explain more of the variability in score differences on the employment interview and the final recommendation than ethnicity. From the fit indices shown in Table 3, it can be concluded that the model fit of the models for the interview and the employment recommendation (χ^2/df of 55.66 and 50.30; TLI of .90; CFI between .94 and .95; RMSEA of .07) was good. Support was found for Hypothesis 2b (see Table 4). For the interview, the explained variance of score differences by Dutch language-proficiency and education was 9% (unstandardized path coefficients of .10 for Dutch language-proficiency [$p < .001$] and .02 for education [$p < .001$]). Ethnicity explained 0.04% (unstandardized path coefficients of .01, $p < .05$) of the variability in test scores. For the employment recommendation, the explained variance of score differences by Dutch language-proficiency and education was 13% (unstandardized path coefficients of .14 for Dutch language-proficiency [$p < .001$] and .02 for education [$p < .001$]). Ethnicity explained 0.09% (unstandardized path coefficients of .02, $p < .05$) of the variability in test scores.

Table 3

Fit Indices of a Cognitive Ability Test (PIT), a Personality Questionnaire (PPV), an AC, an Employment Interview, and a Final Employment Recommendation

Dimension	χ^2	*df*	χ^2/df	TLI	CFI	RMSEA
PIT	5922.79**	106	55.88	.88	.92	.07
PPV	5069.98**	93	54.52	.88	.92	.07
AC	2553.90**	54	47.29	.92	.95	.06
Interview	2449.10**	44	55.66	.90	.95	.07
Final Recommendation	3319.56**	66	50.30	.91	.94	.07

* $p < .05$, ** $p < .001$ (one-tailed).

Table 4

Unstandardized and Standardized Path Coefficients of a Cognitive Ability Test (PIT), a Personality Questionnaire (PPV), an AC, an Employment Interview, and a Final Employment Recommendation

Dimension	Unstandardized Path Coefficients			Standardized Path Coefficients			
	Dutch language-proficiency	Education[a]	Ethnicity[b]	Dutch language-proficiency	Education[a]	Ethnicity[b]	R^2
PIT	.40**	.07**	.10**	.40	.07	.07	.86
PPV	.09**	-.01	-.01	.08	.01	.01	.01
AC	.18**	.01*	.02*	.25	.02	.02	.09
Interview	.10**	.02**	.01*	.29	.06	.02	.12
Final Recommendation	.14**	.02**	.02*	.35	.05	.03	.18

[a] 2 = preparatory vocational education medium level B (VBO-B), 3 = junior general secondary education (MAVO), 4 = senior secondary vocational education (MBO), 5 = high education means senior general secondary education (HAVO), 6 = university preparatory education (VWO), higher professional education (HBO), or research-oriented education (WO).
[b] 1 =Turkish, 2 = Moroccan, 3 = Surinamese, 4 = Antillean, 5 = Western ethnic minorities, 6 = Dutch ethnic majorities.
* $p < .05$, ** $p < .001$ (one-tailed).

Hypothesis 2c predicted that Dutch language-proficiency and education would not explain more of the variability in assessors' evaluations on the AC than ethnicity. Support for this hypothesis was not found. On the AC, as on the interview and the employment recommendation, more variance was explained by Dutch language-proficiency and education (6%) than by ethnicity (0.04%). However, Dutch language-proficiency and education do seem to account for less explained variance on the AC (6%) than on the employment interview (9%) and the final employment recommendation (13%).

Third Group of Hypotheses

Tables 5 and 6 show the results relevant to hypotheses 3 (a-b) which were derived from complexity-extremity theory. No support was found for Hypothesis 3a or 3b. Hypothesis 3a predicted that ethnic majority assessors would rate the ethnic minority group with an excellent Dutch language-proficiency and education higher than the ethnic majority group with the same language-proficiency and education. Score differences on the employment interview and the final recommendation between the ethnic majority group and the ethnic minority group did not exist or were to the advantage of the ethnic majority group. The ethnic minority group members with excellent Dutch language-proficiency and education were rated significantly lower on the employment-recommendation factors of Agency ($t = 2.66$, $p < .05$) and Communion ($t = 2.28$, $p < .05$) than the ethnic majority group with the same Dutch language-proficiency and education.

Hypothesis 3b predicted that ethnic majority assessors would rate the ethnic minority group with low Dutch language-proficiency and education lower than the ethnic majority group with the same language-proficiency and education. The results showed no significant differences between the ethnic majority and minority group with low Dutch language-proficiency and education.

2.4 Discussion

First Group of Hypotheses

Score differences that were found in the literature on the cognitive ability test, the assessment center (AC), and the employment interview, were replicated in the present study by the score differences between the Dutch ethnic majority group and the first-generation minority groups.

Table 5

High Dutch Language-Proficiency and Education: *Means and Standard Deviations of the Majority and Minority Group and their Differences*

	Majority group (N = 400)		Minority group (N = 92)			Levene's test for equality of variances	*t* test for equality of means
Dimension	*M*	*SD*	*M*	*SD*	*d*	*F*	*t*
Interview	4.48	.63	4.38	.60	.16	.06	1.38
Final Recommendation							
Agency	4.41	.82	4.15	.83	.31	.02	2.66*
Communion	4.25	1.06	3.96	1.03	.27	.11	2.28*
Socio-Cultural Awareness	4.41	.62	4.32	.59	.15	.27	1.35

Note. High education means senior general secondary education (HAVO), university preparatory education (VWO), higher professional education (HBO), or research-oriented education (WO). High Dutch language-proficiency means scoring in the top 4%.
d = difference between majority and minority ethnic group means in standard deviation units, positive *d* values indicate higher ethnic majority groups scoring higher.
* $p < .05$, ** $p < .001$ (two-tailed for Levene's test and one-tailed for the *t* test).

Table 6

Low Dutch Language-Proficiency and Education: *Means and Standard Deviations of the Majority and Minority Group and their Differences*

	Majority group (N = 456)		Minority group (N = 173)			Levene's test for equality of variances	*t* test for equality of means
Dimension	*M*	*SD*	*M*	*SD*	*d*	*F*	*t*
Interview	4.04	.51	4.00	.49	.08	.47	.95
Final Recommendation							
Agency	3.62	.77	3.55	.71	.09	.11	.93
Communion	3.66	.94	3.62	.94	.04	.41	.44
Socio-Cultural Awareness	4.06	.50	4.04	.47	.04	.68	.35

Note. Low education means preparatory vocational education medium level B (VBO-B), junior general secondary education (MAVO), or senior secondary vocational education (MBO). Low Dutch language-proficiency means scoring in the lowest 5%.
d = difference between majority and minority ethnic group means in standard deviation units, positive *d* values indicate higher ethnic majority groups scoring higher.
* $p < .05$, ** $p < .001$ (two-tailed for Levene's test and one-tailed for the *t* test).

Noteworthy, striking score differences existed between the first- and the second-generation minority groups. The differences on the personality questionnaire (PPV) were much less systematic with sometimes the majority group and sometimes the minority group scoring higher. Clear systematic differences were found on the dimension Conscientiousness with all ethnic minority groups, both first- and second-generation, scoring higher than the ethnic majority group.

Differences between the first-generation minority group and the second-generation minority group were the largest for the Antillean group, and the smallest for the Turkish group. Turkish minority applicants scored somewhat lower than the other ethnic minority groups on all selection measures. A recent publication by the Dutch National Bureau of Statistics (CBS, 2004) on marks of high school students in The Netherlands reported corresponding results: The Turkish students had poorer results than the Antillean, Moroccan, and Surinamese students. Turkish people have a history of migrant labor. Most of the Turkish people are Muslim and have a strong sense of their own culture and history (e.g., Nijsten, 1998), whereas Antilleans are from Dutch descent. This might be one possible explanation why the differences between the majority group and Turkish minorities remain large, while the differences between the majority group and the second-generation Antillean group is much smaller than the differences between the majority group and the first-generation Antillean group. The Turkish group might be a more separate group because of their strong sense of culture, even after several generations, than the Antilleans who might integrate more easily into Dutch society because of the connection of the Dutch Antilles with The Netherlands. The decrease in Moroccan and Surinamese first- and second-generation minority score differences was in between the decrease from Antillean and Turkish groups. The studies finding consensus on a hierarchy of social distance to the Dutch majority group (e.g., Hraba et al., 1989; Verkuyten et al., 1996) confirm our findings, as these show an ethnic hierarchy where the Antillean minority group is placed on top of the minority groups and the Turkish minority group at the bottom.

The results relevant to Hypothesis 1d showed score differences between first- and second-generation minority groups on all subtests of the cognitive ability test. Score differences were somewhat larger on subtests for crystallized intelligence. These findings are comparable to findings from Bleichrodt and Van den Berg (1995).

Second Group of Hypotheses

Dutch language-proficiency was able to explain more of the variability in

ethnic score differences than education and ethnicity on the cognitive ability test and on the personality questionnaire. However, the results on the personality questionnaire were less profound than the results on the cognitive ability test. Although Dutch language-proficiency did explain more of the variance between test scores on the personality questionnaire than education and ethnicity, in general the entire model did only explain a very small amount of the variance for personality (R^2 was small). Therefore, it can be concluded that other variables than Dutch language-proficiency, education, and ethnicity are possibly related to ethnic score differences on the personality questionnaire. Certain applicant factors may be related to score differences between ethnic groups. One type of applicant factor related to ethnic groups, which Ryan (2001) investigated for cognitive ability tests, is test motivation and test-taking attitudes. These factors, which were not included in the present study, may also influence the scores on personality questionnaires.

Possible explanatory factors for score differences between ethnic groups on subjective measures have had little attention in past research. In this study, explanations were derived from two theories from social psychology, namely assumed-characteristics theory (Locksley et al., 1980; Locksley et al., 1982a; 1982b) and complexity-extremity theory (Linville, 1982; Linville & Jones, 1980). In the present study, these theories were taken out of the lab for the first time. The results from assumed-characteristics theory have demonstrated that knowledge of relevant demographic information diminishes group membership effects. More variance in score differences was explained by Dutch language-proficiency and education on the employment interview and the final recommendation, during which this background information was known, than on the AC, where such knowledge was not given to the assessors. An explanation for the finding that Dutch language-proficiency and education did not explain as much as or less variance than ethnicity on the AC may be that assessors did have some knowledge of the applicants' ethnicity-related demographics just by looking at their behavior and hearing them speak. Research by Jussim et al. (1987, 1996) investigated only one group of assessors, which had knowledge of applicants' demographics. They showed somewhat larger percentages of explained variance (4% for ethnicity, 21% for personal appearance [appearing upper- versus lower-class], and 19% for dialect style [(non-) standard English speaking]) than the results from this study. This was probably due to the highly controlled setting of their lab experiment, explaining why lower percentages of explained variance were found in the present less controlled, but more ecological valid field study.

Third Group of Hypotheses

For complexity-extremity theory, the results were unsupportive. The ethnic

minority and majority groups with a low Dutch language-proficiency and education showed no differences in scores. The ethnic minority group with excellent Dutch language-proficiency and education was not rated significantly higher but, on the contrary, lower on Agency and Communion than the ethnic majority group with the same Dutch language-proficiency and education. Thus, a general tendency seems to exist to rate the ethnic minority group a bit lower than the ethnic majority group on the employment interview and the final employment recommendation. Although systematic and positive *d* values were found, indicating the majority group scoring higher, the effect sizes were very small. Clearly, complexity-extremity processes have not been of influence on assessors' behavior. Tajfel's Social Identity Theory (Tajfel, 1978), which argues that the motivation to maintain a positive social identity and high self-esteem leads to a bias in favor of the in-group, might provide a better explanation for the assessors' evaluations. Maybe other processes, such as demographic similarity between applicants and assessors or perceived similarity of applicants by assessors are responsibly for the score differences that were found in this study. These issues should have more attention in future research.

To summarize this study, three major points are highlighted. Firstly, as expected, score differences between the first- and the second-generation minority groups existed to the advantage of the second-generation minority group. The second-generation minority group did still score lower than the ethnic majority. First- and second-generation minority differences to the advantage of the second-generation minority group existed on both the objective and the subjective measures. They were largest for the Antillean group and smallest for the Turkish group. Secondly, among the ethnicity-related demographic variables Dutch language-proficiency, education, and ethnicity, Dutch language-proficiency and education explained most of the variability in score differences on the employment interview and the final recommendation. This is in line with assumed-characteristics theory. Thirdly, the results were unsupportive for complexity-extremity theory. Other possible explanatory factors for the score differences between ethnic groups on subjective measures, such as demographic and perceived similarity between applicants and assessors, should receive more attention in future research.

Chapter 3

Analyzing judgments of ethnically diverse applicants during personnel selection: A study at the Dutch police [1]

A judgment-analysis study was used to investigate assessors' judgment processes, evaluating ethnic minority versus ethnic majority applicants. Sixteen ethnic majority assessors judged 5,089 applicants during the Dutch police officer selection procedure, with each assessor judging 30 ethnic minority applicants minimally. Information from an employment interview, an assessment center, and a Big Five personality questionnaire were combined into a final employment recommendation. Results showed that as much as or more information sources were used to judge ethnic minority than ethnic majority applicants. Furthermore, a larger number of irrelevant cues were used for the judgment of ethnic minority applicants. Finally, when judging ethnic minority applicants, assessors based their decision to a lesser extent on their own ratings than on ratings of others.

[1] This chapter was published as:
De Meijer, L. A. L., Born, M. Ph., Van Zielst, J., & Van der Molen, H. T. (2007). Analyzing judgments of ethnically diverse applicants during personnel selection: A study at the Dutch police. *International Journal of Selection and Assessment, 15*(2), 139-152. The study in this chapter was also presented at the 21st annual conference of the Society for Industrial and Organizational Psychology (SIOP), Dallas (TX), May 2006.

3.1 Introduction

Human judgment has been studied in a variety of contexts (see Brehmer & Brehmer, 1988, for a review). The purpose of the line of research named *judgment analysis* – previously called *policy capturing* – is capturing the way assessors weigh and integrate information into a statistical model (Hoffman, 1960). This chapter focuses on judgments of assessors in personnel selection. More specifically, we investigate differences that might possibly exist between the judgment processes of assessors judging ethnic minority applicants and of assessors judging ethnic majority applicants. With the term 'judgment process' we mean the process of giving weights to sources of information (e.g., scores on a personality questionnaire and an assessment center exercise) when combining these into a final employment recommendation. Thus, there is no focus on mean subgroup score differences on selection measures but on differences in weights when combining information from various selection measures into a final employment advice.

Although, to our knowledge, the effect of applicant ethnicity on the judgment process has not been investigated until now, a considerable amount of research has examined the interaction between assessor and applicant ethnicity as possible sources of variance in ratings given (McFarland, Ryan, Sacco, & Kriska, 2004). Inconsistent findings have resulted from this body of research. While some studies suggest that assessor by applicant ethnicity interactions do not exist (e.g., Graves & Powell, 1995; Pulakos, White, Oppler, & Borman, 1989; Sacco, Scheu, Ryan, & Schmitt, 2003), others have found they do (e.g., Prewett-Livingston, Field, Veres, & Lewis, 1996). A limitation of such research into interaction effects between assessor and applicant ethnicity is that it does not take into account the judgment process. Pulakos et al. (1989) argued that irrespective of whether there are mean subgroup differences in judgments of assessors, assessors may use different variables or cues in their process of judging someone of a different ethnical background. Judgment analysis is a possible strategy for investigating such similarities and differences in judgment processes.

Firstly, the origin of judgment analysis will be discussed and a description will follow about how human decision processes can be modeled. Secondly, existing research will be highlighted on differences between experienced and inexperienced judges and the effect of experience on the judgment process. This research will be used to form hypotheses on differences in human decision processes when judging ethnic minority versus ethnic majority group members.

Judgment analysis is a methodological application of Social Judgment Theory (SJT) and its underlying framework, Brunswick's Lens Model (e.g., Brunswick, 1952). SJT is used to model the human decision-making process in various types of situations, such as in personnel selection. Despite its name, SJT is not a theory for it provides no testable hypotheses about judgment. Instead, it is a meta-theory that gives direction to research on judgment (Brehmer, 1988). SJT defines judgment as a process, which involves the integration of information from a set of cues into a judgment about certain outcomes (e.g., selection results/outcomes, in case of personnel selection). Usually, for judgment analysis, a statistical model is defined by means of multiple regression analysis. This type of analysis has the ability to express the relationship between the judgment on the one hand, and the weights of the variables or cues used to come to a certain judgment on the other hand, in the form of a linear equation. The resulting regression equation shows the strategy of the assessor and the regression weights reflect the importance of a certain variable or cue awarded by this assessor. Judgment analysis mainly consists of analyses that allow us to identify the weights assigned to pieces of available information during decision-making. This judgment analysis procedure goes back to 1923 (Wallace, 1923). Since then, a large number of other judgment analysis studies have appeared in the literature on clinical diagnosis or on judgment in all sorts of areas, such as in education and in personnel selection (e.g., Barr & Hitt, 1986, McIntyre & James, 1995; Wahlstrom, Hummers-Pradier, Lundborg, Muskova, Lagerlov, et al. 2002).

A review by Graves and Karren (1992) on assessors' decision-making processes suggests a number of possible differences in the judgment formations processes of effective and ineffective assessors. In a number of studies they refer to, the relationship between the effectiveness of assessors and their judgment processes is investigated. A study by Dougherty, Ebert, and Callender (1986, in Graves & Karren, 1992) demonstrated a clear difference between effective an ineffective assessors. After following training in effective assessment, assessors' predictive validities improved. In their review, Graves and Karren (1992) demonstrated that effective assessors based their judgments on less information than ineffective assessors. They also showed that effective assessors used the same sources of information at different times, while ineffective assessors did not necessarily use the same sources. Zedeck, Tziner, and Middlestadt (1983, in Graves & Karren, 1992) also found a similar relationship between effective assessors and use of information. In a study by Kinicki, Lockwood, Hom, and Griffeth (1990, in Graves & Karren, 1992), effectiveness of assessors appeared to be positively related to consistency of their judgments.

Apart from distinguishing between effective and ineffective assessors, a distinction can be made between experienced and inexperienced assessors. In such studies, assessors either are divided in managers and students or assessor-experience is determined more objectively by means of amount of work experience. In agreement with the distinction presented in Shanteau (1991), the following definitions are used in this chapter: 'Experienced assessors' are considered to be the best at what they do, because they have a considerable amount of experience in the field of assessment. 'Inexperienced assessors' may have some experience in the field of assessment but are not yet completely skilled. They are trying to become experienced.

Results from studies in which judgments of experienced and inexperienced assessors were analyzed, showed differences in the use of information. The differences in use of information between experienced and inexperienced assessors appeared largely comparable to the differences between effective and ineffective assessors: Experienced assessors, like effective assessors, make use of less information and of less different sources of information than inexperienced assessors do (Barr & Hitt, 1986; Gorman, Clover, & Doherty, 1978; Singer & Bruhns, 1991). Several other studies, however, found slightly different results regarding the amount of information used by experienced assessors (Ettenson, Shanteau, & Krogstad, 1987; Hammond, Frederick, Robillard, & Victor, 1989; Shanteau, Grier, Johnson, & Berner, 1991). These studies showed that judgments of experienced and inexperienced assessors were based on similar amounts of relevant information. Ettenson et al. (1987) did find that experienced assessors were more consistent in their judgments than inexperienced assessors. Experienced assessors appeared to be better in discriminating between relevant and irrelevant information. These groups of studies show that the judgment processes of effective and ineffective, and experienced and inexperienced assessors can differ from each other. A possible reason why inexperienced and ineffective assessors base their judgments on more and also on more irrelevant information than experienced and effective assessors might be that their lack of experience evokes a need for more information. Furthermore, they possibly cannot differentiate between relevant and irrelevant parts among the various sources of information.

The purpose of the present study is to investigate such differences that might possibly exist between the judgment processes of assessors judging ethnic minority versus ethnic majority applicants. Data of the Dutch police were used, where assessors judged candidates that applied for a training to become a police officer. All assessors at the Dutch police were ethnic majority members who judged far more ethnic majority applicants (the mean number

of ethnic majority applicants per assessor was 280.19) than ethnic minority applicants (the mean number of ethnic minority applicants per assessor was 37.19). An important question then is: When is an assessor experienced in judging ethnic minority and majority applicants? From the so-called contrast model of similarity judgment (Tversky, 1977) it is known that individuals who belong to a certain group (e.g., the ethnic majority group) are familiar with this group, but, in contrast, are not familiar with other groups. Individuals who belong to a certain group have a 'relative richness' of the presentation of the self and others belonging to the same group (Kunda, 1999). As a result, they are experienced within their own group when it comes to judging others in their own group. Applying these findings to ethnic groups, we argue that individuals who are members of the ethnic majority group can be seen as experienced in their own ethnic majority group. These individuals are seen as novices or as inexperienced concerning the ethnic minority group.

In sum, it can be said that assessors at the Dutch police are more experienced in judging ethnic majority applicants than experienced in judging ethnic minority applicants. In line with the research showing that inexperienced assessors use as much or more, and more irrelevant information than experienced judges (e.g., Barr & Hitt, 1986, Singer & Bruhns, 1991), and from the knowledge that assessors have less experience judging ethnic minority applicants, we expect that assessors who judge ethnic minority applicants will use as much or more, and more irrelevant information than assessors who judge ethnic majority applicants. From this, the following hypothesis was derived.

> *Hypothesis 1:* Ethnic majority assessors judging ethnic minority applicants will use as much or more cues than assessors judging ethnic majority applicants.

Following from the evidence that inexperienced assessors use more irrelevant information to come to their final judgment than experienced assessors do (Ettenson et al., 1987), it is expected that assessors who judge ethnic minority applicants will use more irrelevant information than assessors who judge ethnic majority applicants.

> *Hypothesis 2:* Ethnic majority assessors judging ethnic minority applicants will use more irrelevant cues than assessors judging ethnic majority applicants.

A final issue that we investigate is whether assessors use their own ratings or ratings of others (i.e., from other assessors or from the applicant) as sources in their judgments. As mentioned earlier, a possible reason why inexperienced assessors base their judgments on more and also on more irrelevant

information than experienced assessors (e.g., Barr & Hitt, 1986, Singer & Bruhns, 1991) might be that their lack of experience evokes a need for more information. Because of their lack of experience, it is likely that inexperienced assessors are more uncertain in the decision-making process. It is expected that, as a result of their uncertainty, they base their decisions to a lesser extent on their own judgments. Consequently, inexperienced assessors may use ratings of others (i.e., from other assessors or from the applicant) as sources in their decision-making process.

During the selection procedure at the Dutch police, the psychologist who conducts the interview is the one who also writes the final recommendation. The employment interview is, therefore, a source of information coming from the assessor him- or herself, while the AC- and personality questionnaire factors are cues coming from others, that is from another assessor and from the applicant.

> *Hypothesis 3:* Ethnic majority assessors judging ethnic minority applicants will base their decision to a lesser extent on their own ratings than on ratings of others than when judging ethnic majority applicants.

3.2 Method

Firstly, two methodological issues need to be discussed using judgment analysis as an approach to clarify the judgment strategies used by assessors. The first issue concerns the use of linear regression, which assumes that the relation between each cue and the judgment is linear. If the linearity assumption is unreasonable then the linear model may be misleading. The linear model should be abandoned reluctantly, however, for to do so may introduce complexities into the analysis that outweigh possible gains in accuracy. The linear model has the advantage that it can accurately describe many processes that are not strictly linear (Dawes & Corrigan, 1974). Furthermore, results from research with linear models have been reviewed many times (Brehmer, 1994). Brehmer and Brehmer (1988) for instance found that linear models fit judgments quite well. When configural components were found, they usually accounted for only a few percent of the variance. Because of the above-mentioned advantages of a linear model, we used linear regression to examine judgment processes.

The second methodological issue in analyzing judgment strategies of individual assessors is the choice of an experimental or non-experimental

design. In experimental judgment-analysis studies, manipulating a number of variables or cues in a balanced factorial design creates profiles of hypothetical applicants. The key advantage of this approach is that the correlations between the cues are zero. The key disadvantage of experimental designs is the use of hypothetical applicants. Hypothetical profiles may not offer a representative simulation of real profiles and may lack external validity (Gorman et al., 1978; Hobson & Gibson, 1983; Karren & Barringer, 2002). In non-experimental designs, assessors generally evaluate real applicants on a number of predetermined selection variables or cues and make an overall judgment. The primary advantage of these designs is that researchers study assessors' evaluations of real applicants. The external validity problems created by the use of hypothetical applicants are not present. The primary disadvantage of non-experimental designs is that the selection variables might be correlated. As a result it is often difficult to determine the relative importance of the selection variables to each assessor's decisions (Karren & Barringer, 2002). Assessors' strategies for integrating the variables or cues to make judgments then cannot readily be identified. Research in which estimates of cue importance under three different correlation structures where compared (Lane, Murphy, & Marques, 1982), nevertheless, found that zero intercorrelations are nót required to estimate the importance of explanatory variables. Cue intercorrelations should, however, be minimized.

While the use of hypothetical applicants may be useful in experimental designs, Gorman et al. (1978) noted many weaknesses in such an approach. They argued, for instance, that assessors cannot view the applicants, while they normally can view them in the situation of an employment interview. Gorman et al. (1978) conducted two investigations of the validity when using hypothetical job applicants and concluded that substantially more valid decisions were made when the assessor actually viewed true applicants. In line with the results, to obtain accurate data, we used actual judgment data. Therefore, the present study is a field study and analyzes judgments from actual assessors rating actual applicants. We take into consideration that cue intercorrelations should be minimized.

Participants and Procedure

Data came from 16 assessors from the Police Academy of The Netherlands, who evaluated 5,089 applicants from September 2001 until July 2003. Each assessor evaluated a minimum of 30 ethnic minority applicants. All assessors were psychologists and ethnic majority group members. Two assessors (13%) were male and fourteen (87%) were female. Furthermore, 35% of the assessors were 30 years old or older, and 62% were younger than 30. From the 5,089 applicants who were evaluated, 270 applicants (5%) were first-

generation ethnic minorities (minority group members who had moved to The Netherlands after they were born) and 325 applicants (6%) were second-generation ethnic minorities (minority group members who were born in The Netherlands but at least one of their parents was born outside The Netherlands). The professions for which accepted applicants were to be trained were assistant police employee, police employee, or all-round police employee. Applicants who were interested in a job as police officer first applied to the local police force where they wanted to work after they would complete their training. For the selection procedure, the local police forces routinely send all applicants to the national Police Center for Competence Assessment and Monitoring (CCM). During the selection procedure of the CCM, several selection measures are used. The present study focused on the following measures, namely: the assessment center (AC), the employment interview, and the personality questionnaire. The psychologist who conducted the interview was also the one who wrote the final recommendation to the police force. For the final recommendation, the test results of the personality questionnaire, the AC ratings, and the employment interview ratings were used (for a schematic presentation of the selection procedure as a whole, see Appendix A).

Table 1 shows the characteristics of the groups of applicants in terms of demographic variables. This study focused on the Dutch majority groups and the largest first- and second-generation ethnic minority groups in The Netherlands, which are the Dutch Antillean, the Moroccan, the Surinamese, and the Turkish group. Twelve percent of the applicants were ethnic minorities and 88% were ethnic majorities. The percentage male of the first-generation minority, the second-generation minority, and the majority group was, respectively, 72%, 71%, and 65%. The mean age of the first-generation minority, the second-generation minority, and the majority applicants was, respectively, 27, 22, and 24 years old.

Measures

Personality

To measure the Big Five factors Extraversion, Altruism, Conscientiousness, Emotional Stability, and Intellect, the Police Personality Questionnaire (PPV; Van Leeuwen, 2000) was used. The applicants completed the PPV in Dutch. A recent progress report by Klinkenberg and Van Leeuwen (2003) indicated alpha reliabilities (N = 5,641) of .76 for Extraversion (sample item: "Social contact is important to me", .75 for Altruism (sample item: "I like to work with other people"), .72 for Conscientiousness (sample item: "I like to work in a structured way"), .74 for Emotional Stability (sample item: "I worry about things"), and .78 for Intellect (sample item: "I am prepared to take a different

point of view than other people"). Each scale consists of 10 items. Factor analysis, using the scree-plot criterion, yielded a five-factor structure (the five factors explained 39% of the variance; the mean factor loadings per factor varied between .53 and .57 [N = 6,226]). Correlations between the scales are all lower than .60. Comparison with NEO-PI-R showed observed construct validity coefficients of .58 (p < .05) for Extraversion, .34 (p < .05) for Altruism, .47 (p < .05) for Conscientiousness, .59 (p < .05) for Emotional Stability, and .17 (p < .05) for Intellect (N = 160). A study by Lem and Van Doorn (2000) showed observed validity coefficients between .15 and .43 (N = 61) for the prediction of supervisory evaluations of job performance.

Table 1

Distribution of Majority Group Members and First- and Second-Generation Minority Group Members in Terms of Gender and Age

	n (%)	% male	M_{age} (*SD*)
1st Generation minority group	270 (5.3)	72	27.27 (7.09)
2nd Generation minority group	325 (6.4)	71	21.87 (5.11)
Majority group	4,483 (88.1)	65	23.98 (7.07)
Total	5,089 (100)	66	24.03 (7.03)

Note. Of 11 applicants it was not known if they were first-generation ethnic minority, second-generation ethnic minority, or majority group members.

Assessment Center (AC) and Employment Interview

Article 2 of the Dutch police law states: "The police force has the duty of ensuring the effective maintenance of the legal order and helping those in need." (cf. Van Loon, 2003). This definition reflects the core of the police task. Based on this article and a thorough job analysis conducted by psychologists at the Dutch police who are experienced in job analysis design and administration, an assessment center (AC) and an employment interview have been developed to measure the following twelve dimensions: Communication Skills, Social Skills, Empathy, Initiative, Flexibility, Stress Tolerance, Emotional Stability, Authority, Decisiveness, Tolerance Towards Others, Integrity, and Self-Understanding. For an overview of these dimensions and their definitions, see Appendix B.

Assessment center (AC). The AC is designed to evoke authentic behavior of applicants. Most applicants have no relevant police work experience. Therefore, the AC – which means, a role-play exercise in the case of the Dutch police selection – is not a direct simulation of police work. Three people participate in each role-play, namely: An actor, an assessor, and an applicant. All actors and assessors, who received a higher professional education ("HBO") or research-oriented education ("WO"), are formally trained during a three-week training period as an actor as well as an assessor in role-play exercises. The assessors and the actors, alternately, act and assess in subsequent role-plays.

The following procedure is used during the role-play exercise. Preparation of the role-play begins when the assessor guides the applicant to a room where (s)he has 15 minutes to read a written instruction. The assessor does not provide any further information. Applicants go through the tutorial that teaches them about the fictitious situation, including information about the role of the applicant and the fictitious things that happened before (e.g., being a floor-manager of an airline). After 15 minutes, during which the applicant studies the instructions, the assessor guides the applicant to the door of a room. The only two things the assessor says is: „Do you have any questions so far?" and „The simulations starts as soon as you enter the room." The assessor enters the room, leaving the applicant outside. Inside the room, there is an actor. Actual assessment begins when the applicant opens the door. The actor follows a detailed script on how to interact with the applicants. The assessor, who is not role-playing, listens to the applicant and takes notes of what the applicant says and does.

At the end of the role-play exercise, which takes 15 minutes, the actor and the assessor in the role-play independently make ratings on a 7-point Likert-scale ranging from 1 (extremely weak) to 7 (excellent), on each of the following seven dimensions: Communication Skills, Social Skills, Empathy, Initiative, Stress Tolerance, Authority, and Decisiveness. After the actor and the assessor have completed their independent ratings, they discuss each applicant to reach consensus on the final dimension ratings. Here, they also use a 7-point scale. Interrater reliabilities of the independent ratings of actors and the assessors, given prior to the moment consensus was reached, ranged from .82 to .88 (established from a sample of: N_{actor} = 198 and $N_{assessor}$ = 198). Principal component analysis with varimax rotation yielded two factors, Agency and Communion (in accordance with Wiggins and Trapnell, 1996), which explained 77% of the variance. As a measure of Agency, the average rating across the dimensions Authority, Decisiveness, Initiative,

Communication Skills, and Stress Tolerance was used ($\bar{r}$ = .59; α = .87). As a measure of Communion, the average rating of the dimensions Social Skills and Empathy was used ($\bar{r}$ = .77; α = .87). The reliability of the difference (r_{diff}) between scores on Agency and Communion is .78.

Employment interview. The interview questions are focused on evaluating behavior on the following eight dimensions: Communication Skills, Social Skills, Flexibility, Stress Tolerance, Emotional Stability, Tolerance Towards Others, Integrity, and Self-Understanding.

A single interviewer conducts the interview. The interviewers have received a research-oriented education ("WO") and they are formally trained during a four-week training period. The interviews are semi-structured and behaviorally based (cf. Janz, 1982), with one behaviorally anchored 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent) for each of the eight dimensions. A semi-structured interview combines a highly structured agenda using fixed questions with the flexibility to ask additional questions. In a behaviorally based interview, a candidate is asked to pinpoint specific instances in which particular behavior was exhibited in the past. The general idea is that behavior exhibited in the past is predictive for behavior that will be exhibited in the future. For each dimension, the interviewer chooses an initial question from a list. Sample questions for, e.g., Emotional Stability are: "Can you tell me something about a specific difficult period in your life?", "What was the impact of it, during that time?", "How did you cope with it?", and "How do you deal with it, at present?" The interviewer is instructed to ask additional questions until the dimension can be comprehensively evaluated. The interviewer takes notes during the interview regarding the applicant's reported behaviors. After the interview is completed, the interviewer reviews his or her notes and rates each dimension.

Data from 16 interviewers were used in the present study. A mean number of 317.38 applicants were interviewed per interviewer. The mean number of ethnic majority applicants per interviewer was 280.19 and the mean number of ethnic minority applicants per interviewer was 37.19. The average rating across the eight dimensions was used as the dependent variable as the ratings were substantially correlated ($\bar{r}$ = .42; α = .85). Moreover, principal component analysis with varimax rotation yielded one general interview factor that explained 50% of the variance.

Final Employment Recommendation

The final decision as to whether an applicant is fit for a job as police officer,

is based on several ratings. The psychologist who conducts the interview also gives the final recommendation. The psychologist makes use of the results of the personality questionnaire (PPV), the AC, and the employment interview and integrates all scores into a final recommendation to the local police force where the applicant first applied. The eleven dimensions in the final recommendation are: Communication Skills, Social Skills, Empathy, Initiative, Flexibility, Stress Tolerance, Authority, Decisiveness, Tolerance Towards Others, Integrity, and Self-Understanding. A 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent) is used to evaluate the behavior on the eleven dimensions. Principal component analysis with varimax rotation yielded three final recommendation factors, Agency, Communion, and Socio-Cultural Awareness, which altogether explained 67% of the variance. As a measure of Agency, the average rating across the dimensions Authority, Decisiveness, Initiative, Communication Skills, Stress Tolerance, and Flexibility was used ($\bar{r}$ = .48; α = .85). As a measure of Communion, the dimensions Social Skills and Empathy were used ($\bar{r}$ = .66; α = .79), and for Socio-Cultural Awareness the dimensions Tolerance Towards Others, Integrity, and Self-Understanding ($\bar{r}$ = .39; α = .65). The reliability of the difference (r_{diff}) between scores on Agency and Communion is .51, r_{diff} between scores on Agency and Socio-Cultural Awareness is .58, and r_{diff} between scores on Communion and Socio-Cultural Awareness is .57.

Analyses

Preliminary Analyses

Because response styles can affect answers on questionnaires (e.g., Van Herk, Poortinga, & Verhallen, 2004), structural equivalence (i.e., absence of bias) of the personality questionnaire (PPV), the AC, the employment interview, and the final recommendation were checked before conducting further analyses. In accordance with Van Herk et al. (2004), *structural equivalence* is interpreted as follows: A test measures the same trait cross-culturally, but not necessarily on the same quantitative scale. Using Amos 5.0 (Arbuckle, 2003), no differences between factor structures of all selection measures and the final employment recommendation were found between the majority group and the minority group.

Although all assessors in the present study were formally trained and had a considerable amount of experience in the field of personnel assessment, we wanted to check whether sub-group differences existed in overall experience (i.e., the total number of applicants assessors have assessed) between assessors. Because possible differences in overall experience may have a contaminating effect on the subject studied here – namely the differences

between judgments given to ethnic majority versus ethnic minority applicants – multiple-group analysis was used to look into differences in cue use between more overall-experienced and less overall-experienced assessors. Assessors at the Dutch police are regarded as less experienced during the first 6 months of their employment. During this period of employment, they have no prior experience as assessor and they are supervised. After these first six months, assessors work independently and without supervision. Multiple-group analysis using Amos 5.0 (Arbuckle, 2003) showed no differences in cue use between more overall-experienced and less overall-experienced assessors. This means that the assessors can be viewed as equally experienced in assessing applicants in general. For detailed information on the structural equivalence and the multi-group analysis, the first author may be contacted.

Main Analyses

Multiple regression analysis develops an equation to express the relation between one variable, called the dependent variable, and several others, called the predictors or the independent variables. In the case of judgment analysis, the dependent variable is the judgment and the independent variables are the cues. Regression analysis is used to determine weights for the cues as an estimate of their importance for the judgments. The regression equation can be used to predict an individual assessor's judgment. The accuracy of such predictions depends on how well the regression model fits the assessor's policy and how consistently the assessor applies the policy. A high value of R^2 indicates that the model fits well and that the assessor is highly consistent (Stewart, 1988). The cue weights, derived from multiple regression analysis, are unambiguous only if intercorrelations of cues are low (Stewart, 1988).

Standardized regression weights or β weights are the weights obtained from regression analysis involving the cue importance expressed in standard score form. The relative magnitudes of the standardized β weights for different cues can be directly compared. Standardized β weights are also generally superior to cue-intercorrelations, because the procedure for deriving β weights controls for variation on other variables. Stewart (1988) argued that the β weight for a cue provides an estimate of its effect on judgment with the other cues held constant. Therefore, moderate intercorrelations between cues are acceptable. Cues in the present field study had a moderate mean intercorrelation (r_{mean} = .25, varying from .01 to .63; r_{median} = .25). Although Lane et al. (1982) and Stewart (1988) found that zero intercorrelations are not required to estimate the importance of explanatory variables, we want to highlight that only three, out of 28, intercorrelation were high (r > .50).

Structural equation modeling with Amos 5.0 (Arbuckle, 2003) was used to

investigate the relation between the various cues or independent variables and the final employment recommendation or dependent variables. Because some cues were intercorrelated, a model was hypothesized where cue-intercorrelations were taken into account and, therefore, controlled for. Multi-group analysis was used to look into the differences between ethnic minority and majority applicants. Beta weights for the ethnic majority and the ethnic minority group could, therefore, be directly compared.

3.3 Results

In general, the linear model predicted the observed data quite well. Fit indices and squared multiple correlations (R^2) of the aggregated models of all 16 psychologists together for the final recommendation factors Agency, Communion and Socio-Cultural Awareness are reported in Table 2. The following fit indices were chosen in order to get an impression of the overall fit of the various models: 1) the incremental fit index (IFI [Bollen, 1989]); 2) the comparative fit index (CFI [Bentler, 1990]); and 3) the root mean squared error of approximation (RMSEA [Hu & Bentler, 1995]). IFI and CFI values close to 1 indicate a very good fit. RMSEA values of about .08 or less indicate a close fit of the model.

Firstly, the model fit of the aggregated models (i.e., of the 16 psychologists combined) is mentioned. Multi-group analysis is used to investigate overall differences in the de decision-making process of assessors judging ethnic majority and ethnic minority applicants. Secondly, results testing the three hypotheses will be reported. Addressing the hypotheses, the decision-making processes of 16 psychologists are looked into separately.

The overall fit of the aggregated models was good for all final recommendation factors. The R^2 of the aggregated models varied between .69 and .93. The R^2 for the final recommendation factor Agency was .93 for ethnic minority applicants and .91 for ethnic majority applicants. For the factor Communion overall R^2 was .83 for the minority applicants as well as the majority applicants, and for Socio-Cultural Awareness R^2 was .73 for the minority applicants and .69 for the majority applicants. As can be seen from these results, the differences between explained variance by the model differed only marginally between judgments of ethnic minority and ethnic majority applicants.

Table 2

Fit Indices and Squared Multiple Correlations for the Aggregated Models of the Final Recommendation Factors Agency, Communion, and Socio-Cultural Awareness

Final Recomm. Factor	Fit Indices				R^2
	χ^2 (*df* = 34)	IFI	CFI	RMSEA	
Agency	1188.30**	0.95	0.95	0.08	**0.93**/0.91
Communion	1188.64**	0.94	0.94	0.08	**0.83**/0.83
Soc.-Cult. Aw.	1189.11**	0.93	0.93	0.08	**0.73**/0.69

Note. Coefficients for ethnic minority applicants are in **bold.**
** $p < .001$.

To determine whether differences in judgments in general existed when evaluating ethnic minority versus and ethnic majority applicants, multi-group analysis was used. Differences in β weights were analyzed using Amos 5.0 (Arbuckle, 2003). The multi-group analysis was conducted on two equations (i.e., two assessor-specific judgments): One equation for the judgment of ethnic minority applicants and one equation for the judgment of ethnic majority applicants. Firstly, the multi-group analyses were conducted for the aggregated models, i.e., on each of the final recommendation factors for all 16 psychologists combined. Secondly, differences were examined on the individual level, i.e., for each of the 16 psychologists. Results of the aggregated multi-group analysis yielded the following results: 1) for the final recommendation factors Agency ($\Delta\chi^2(\Delta df = 8) = 16.55$, $p < .05$) and Communion ($\Delta\chi^2(\Delta df = 8) = 16.21$, $p < .05$), significant differences existed between the cue-usage of judgment of ethnic minority and majority applicants; 2) for the final recommendation factor Socio-Cultural Awareness ($\Delta\chi^2(\Delta df = 8) = 14.23$, *ns*) no significant differences existed. Because no overall differences existed in the use of cues by assessors when judging ethnic minority versus ethnic majority applicants on the latter factor, this factor was omitted from the multi-level analyses for each of the 16 psychologists separately.

For the separate 16 psychologists, addressing hypotheses 1 to 3, the β weights and the differences in chi-square for the final recommendation factors Agency and Communion are shown in Tables 3 and 4. Standardized β weights

for the judgment of ethnic minority applicants are in italic. Significant differences between β weights are in bold.

We expected that assessors judging ethnic minority applicants used as much or more cues than assessors judging ethnic majority applicants (Hypothesis 1). For the final recommendation factor Agency as well as Communion, nine of sixteen psychologists (56%) weighted cues differently when judging ethnic majority versus ethnic minority applicants. Taking the final recommendation factors Agency and Communion together, 256 comparisons were made (2 final recommendation factors x [16 psychologists x 8 cues]). Twenty-six of these 256 comparisons in β weights (10%) were significantly different. On the basis of coincidence, 5% of the differences ($\alpha = .05$) would be significant. Hence, it can be concluded that these 26 differences are actual differences. For eleven significant differences between standardized β weights (of 26 significant differences), the β weight differed significantly from zero for the judgment of ethnic minority applicants but did not differ significantly from zero for the judgment of ethnic majority applicants. It can therefore be concluded that cues of these eleven differences were used for the judgment of ethnic minority applicants and were not used for the judgment of ethnic majority applicants. For eight differences between cues (of 26), a significantly larger weight was allocated for the judgment of ethnic minority applicants than for the judgment of majority applicants. For another seven differences between cues (of 26), a significantly larger weight was allocated for the judgment of ethnic majority applicants. These findings were in contradiction with what we expected. However, 97% (249 of 256) of the cues were used as much as or more by assessors for the judgment of ethnic minority applicants, which supports findings from Barr and Hitt (1986) and Singer and Bruhns (1991). Therefore, it can be concluded that the results concerning the final recommendation factors Agency and Communion largely support Hypothesis 1. Previously, we found that for the final recommendation factor Socio-Cultural Awareness the same amount of cues was used for the judgment of ethnic minority versus majority applicants. These results are also supportive of Hypothesis 1. In the Discussion we will return to the final recommendation factor Socio-Cultural Awareness and the difference in its results compared to Agency and Communion.

Hypothesis 2 expected that assessors judging ethnic minority applicants used more irrelevant cues than assessors judging ethnic majority applicants. The correct utilization of the cues was established by looking at intercorrelations between a final recommendation factor (Agency or Communion) and the various cues (for definitions of dimensions which are clustered in terms of Agency and Communion, see Method section and Appendix B).

Table 3

Relations Between the AC, the Employment Interview, and the Personality Questionnaire (PPV) on the one hand and on the other hand the Final Recommendation Factor Agency for Ethnic Minority and Majority Applicants

Psychologist[a]			Standardized β Weights[b] $\Delta\chi^2(\Delta df = 1)$[c]							
	N	R^2	AC Agency	AC Comm.	Interview	Extrav.	Altruism	Consc.	Emot. St.	Intellect
1	*35*/182	*.97*/.93	*.75*/.73	*-.10*/-.09	*.37*/.37	*.02*/ .01	*-.04*/-.05	*.02*/-.02	*-.02*/ .04	*.07*/.04
			0.17	0.12	1.37	0.08	0.01	0.46	1.33	0.68
2	*36*/270	*.96*/.93	*.80*/.73	*-.12*/-.11	*.29*/.42	*.07*/ .02	*-.07*/-.04	*-.03*/ .00	*.04*/ .03	*-.02*/.04
			3.65	0.23	1.71	0.79	0.39	0.46	0.05	0.56
3	*35*/261	*.93*/.88	*.71*/.66	*-.12*/-.02	*.42*/.42	*.16*/ .04	*-.14*/-.02	*.00*/ .03	*-.05*/ .04	*.05*/.07
			1.53	1.83	0.15	2.96	2.94	0.11	1.33	0.08
4	*38*/345	*.95*/.93	*.74*/.77	*.02*/-.03	*.30*/.38	*-.02*/ .01	*.03*/-.02	*.00*/-.02	*.01*/ .02	*.00*/.02
			0.38	0.59	1.58	0.22	0.45	0.04	0.03	0.06
5	*55*/378	*.96*/.91	*.69*/.76	*-.06*/-.08	*.43*/.37	*.11*/ .02	*-.01*/-.02	***-.13*/ .00**	*.02*/ .03	*.05*/.04
			1.53	0.12	2.89	3.79	0.02	**10.28***	0.03	0.09
6	*30*/144	*.93*/.89	*.67*/.69	*-.03*/-.09	*.46*/.41	*-.04*/ .10	*.07*/-.04	*-.02*/ .05	*.00*/-.01	*.02*/.05
			0.91	0.40	0.06	2.67	1.25	1.03	0.06	0.12
7	*30*/204	*.98*/.93	*.76*/.76	*.03*/-.01	*.28*/.34	*.00*/ .01	***-.14*/ .00**	*.02*/ .00	*.03*/ .03	*.05*/.03
			0.34	0.77	0.07	0.05	**8.73***	0.19	0.03	0.30
8	*36*/413	*.94*/.92	*.70*/.61	*-.12*/-.03	*.46*/.50	*-.01*/ .03	*-.10*/-.04	*-.01*/-.02	*-.05*/ .04	*.06*/.03
			0.49	1.60	2.02	0.73	0.88	0.06	2.38	0.16
9	*40*/257	*.93*/.92	*.73*/.84	*.06*/-.05	*.24*/.24	***.23*/ .06**	*-.14*/-.03	*-.02*/ .03	*-.01*/ .05	*-.01*/.02
			3.28	2.72	2.80	**5.54***	1.94	0.61	1.57	0.24

10	*61*/398	*.93*/.92	*.83*/.78	*-.10*/-.07	***.28*/.35**	*.03*/ .02	*-.01*/-.03	*-.01*/-.01	*.06*/ .02	*.04*/.04
			1.23	0.45	**8.35***	0.05	0.13	0.00	0.12	0.01
11	*32*/304	*.93*/.93	*.54*/.68	*-.10*/-.04	*.57*/.43	*-.05*/ .00	*-.09*/-.03	*-.04*/ .00	*.04*/ .02	*.04*/.04
			0.91	0.53	2.53	0.58	0.48	0.38	0.15	0.00
12	*30*/282	*.96*/.94	*.86*/.80	*.02*/-.05	***.31*/.33**	*.03*/ .02	*.02*/-.06	*-.04*/ .01	*.01*/ .02	*.06*/.03
			0.10	2.92	**3.88***	0.00	2.46	0.54	0.04	0.08
13	*42*/308	*.96*/.93	*.72*/.80	*-.06*/-.04	*.35*/.26	***.17*/-.01**	*-.01*/-.02	*-.03*/-.01	*-.06*/ .02	***-.09*/.04**
			1.60	0.20	3.11	**7.43***	0.03	0.45	2.92	**5.68***
14	*31*/267	*.94*/.90	*.74*/.81	*.00*/-.07	*.35*/.28	*.13*/ .01	*.03*/-.03	*-.01*/ .03	*.03*/ .01	***-.10*/.06**
			0.53	1.09	0.13	2.95	0.91	0.30	0.16	**4.97***
15	*31*/230	*.92*/.87	*.66*/.50	***-.38*/-.14**	***.63*/.59**	*.07*/ .01	***-.31*/-.04**	*-.09*/ .01	*.10*/ .06	***.21*/.00**
			1.42	**4.88***	**9.34***	0.41	**5.39***	1.26	0.08	**5.24***
16	*33*/240	*.96*/.91	***.84*/.73**	*-.13*/-.04	*.31*/.37	*-.03*/ .00	*-.01*/-.04	*-.01*/-.01	*.06*/ .03	*.11*/.05
			8.67*	3.13	0.51	0.30	0.11	0.01	0.50	0.65

Note. AC Comm. is AC factor Communion, Extrav. is Extraversion, Altruism is Altruism, Consc. is Conscientiousness, Emot. St. is Emotional Stability, and Intellect is Intellect.

[a] Sixteen psychologists were selected, who evaluated at least 30 ethnic minority applicants each.

[b] Coefficients for ethnic minority applicants are in *italic*. Significant differences between β weights are in **bold**.

[c] Differences in chi-square are given for the multi-group analysis (* $p < .05$).

Table 4

Relations Between the AC, the Employment Interview, and the Personality Questionnaire (PPV) on the one hand and on the other hand the Final Recommendation Factor Communion for Ethnic Minority and Majority Applicants

Psychologist[a]			Standardized β Weights[b] $\Delta\chi^2(\Delta df = 1)$[c]							
	N	R^2	AC Agency	AC Comm.	Interview	Extrav.	Altruism	Consc.	Emot. St.	Intellect
1	*35*/182	*.90*/.89	*-.07*/-.09	*.64*/.70	*.43*/.40	*.15*/.09	*-.17*/ .05	*.07*/ .00	*.06*/-.04	*-.06*/-.07
			0.08	1.06	0.11	0.30	3.82	0.64	2.00	0.10
2	*36*/270	*.89*/.89	*-.09*/-.05	*.95*/.87	*.03*/.18	*.09*/.08	***.24/ .01***	*.02*/ .02	*-.03*/-.06	*-.14*/-.01
			0.10	0.00	2.84	0.00	**5.37***	0.00	0.57	0.95
3	*35*/261	*.87*/.80	***-.44/-.04***	*.65*/.67	***.79/.42***	*.08*/.19	*.06*/ .06	*.09*/ .01	*-.05*/-.10	*.00*/-.08
			7.49*	0.52	**7.23***	1.25	0.00	0.76	0.23	0.91
4	*38*/345	*.94*/.88	*.03*/ .00	*.86*/.83	***.16/.29***	*.00*/.02	*.02*/ .04	*-.08*/-.03	*.00*/-.05	*.02*/ .02
			0.11	1.52	**7.69***	0.09	0.06	0.85	0.55	0.00
5	*55*/378	*.85*/.89	*.01*/-.07	*.69*/.73	*.38*/.39	*.17*/.10	*.00*/ .02	*-.01*/ .02	*.00*/-.02	*-.11*/-.07
			0.93	0.47	0.63	0.69	0.04	0.18	0.13	0.25
6	*30*/144	*.86*/.84	*-.10*/-.08	*.75*/.71	*.39*/.36	*.10*/.10	*.08*/-.08	*-.04*/ .01	*.02*/-.01	*-.20*/-.03
			0.05	0.95	0.13	0.00	1.61	0.18	0.11	2.08
7	*30*/204	*.91*/.88	*.06*/-.07	*.84*/.84	*.15*/.32	*-.09*/.02	*-.07*/ .03	*.01*/ .00	*.08*/ .00	*.02*/-.01
			1.37	1.17	2.17	1.04	1.40	0.02	0.99	0.13
8	*36*/413	*.93*/.83	*-.05*/-.03	***.84/.75***	*.28*/.35	*.03*/.04	*.09*/ .04	*-.09*/ .01	***.12/-.06***	*-.10*/-.03
			0.17	**4.85***	0.03	0.02	0.70	2.21	**6.09***	1.60
9	*40*/257	*.92*/.92	*.10*/-.03	*.76*/.88	*.19*/.18	*-.04*/.06	***.19/ .03***	*-.03*/ .01	*.12*/ .02	*-.05*/-.05
			2.92	1.44	0.87	1.88	**4.58***	0.33	3.14	0.01

10	*61*/398	*.87*/.82	*-.08*/-.10	*.81*/.80	***.32*/.35**	*-.03*/.03	*-.05*/-.02	*-.08*/ .03	*.10*/ .01	***.09*/-.05**
			0.07	0.23	**5.11***	0.86	0.29	3.32	2.49	**5.94***
11	*32*/304	*.88*/.81	*.04*/-.17	*.73*/.71	***.23*/.45**	*.01*/.06	*-.02*/ .00	*-.12*/ .01	*.09*/-.06	*.07*/-.06
			2.58	0.22	**4.54***	0.45	0.04	2.35	3.00	1.95
12	*30*/282	*.91*/.78	*.02*/-.17	***.94*/.80**	***.12*/.37**	*.12*/.05	*-.13*/ .03	*-.03*/ .01	*-.07*/-.08	*.06*/-.03
			3.45	**3.91***	**6.83***	0.52	2.45	0.11	0.00	1.00
13	*42*/308	*.87*/.89	*-.05*/-.02	*.82*/.86	*.21*/.17	*.12*/.01	*-.08*/ .04	*-.01*/-.02	*-.09*/-.01	*-.06*/ .02
			0.06	0.00	0.23	1.07	0.34	0.00	0.89	0.83
14	*31*/267	*.91*/.82	*-.21*/-.13	*.93*/.83	*.28*/.29	***.21*/.01**	*.02*/ .08	*.00*/ .01	*-.09*/-.04	*-.16*/-.08
			1.32	1.33	0.04	**4.88***	0.40	0.02	0.71	0.78
15	*31*/230	*.80*/.67	*.08*/-.08	*.46*/.65	*.43*/.36	*.14*/.09	*.01*/-.06	*.15*/ .02	*-.11*/-.04	*-.23*/-.02
			0.67	0.10	0.59	0.26	0.09	1.17	0.45	3.23
16	*33*/240	*.79*/.77	*-.17*/-.03	*.86*/.71	*.25*/.29	*-.06*/.11	*.16*/ .08	*-.04*/-.04	*.01*/ .01	*.02*/-.08
			0.92	0.00	0.52	3.01	0.25	0.00	0.00	0.50

Note. AC Comm. is AC factor Communion, Extrav. is Extraversion, Altruism is Altruism, Consc. is Conscientiousness, Emot. St. is Emotional Stability, and Intellect is Intellect.

[a] Sixteen psychologists were selected, who evaluated at least 30 ethnic minority applicants each.

[b] Coefficients for ethnic minority applicants are in *italic*. Significant differences between β weights are in **bold**.

[c] Differences in chi-square are given for the multi-group analysis (* $p < .05$).

Relevant cues to come to a final recommendation on Agency are the AC factor Agency, the interview, and the personality questionnaire factors Extraversion, Emotional Stability, and Intellect. Irrelevant cues for a final recommendation on Agency are the AC factor Communion and the personality questionnaire factors Altruism and Conscientiousness. Relevant cues to come to the final recommendation factor Communion are the AC factor Communion, the interview, and the personality questionnaire factors Extraversion, Altruism, and Intellect. Irrelevant cues are the following three: the AC factor Agency and the personality questionnaire factors Conscientiousness and Emotional Stability. For the examination of Hypothesis 2, significant differences in β weights of irrelevant cues were checked for the final recommendation factors Agency and Communion. For the final recommendation factor Agency, three of sixteen psychologists (19%) used more irrelevant cues or gave more weight to irrelevant cues when they judged ethnic minority applicants than when they judged ethnic majority applicants. For the final recommendation factor Communion, this was the case for two of sixteen psychologists (13%). For four cues, the standardized β weights differed significantly from each other, differing significantly from zero for the judgment of ethnic minority applicants but not significantly from zero for the judgment of ethnic majority applicants. It can be concluded that these four cues were used for the judgment of ethnic minority applicants but not for the judgment of ethnic majority applicants. For two cues, a significantly larger weight was allocated for the judgment of ethnic minority applicants than for the judgment of majority applicants. These results support Hypothesis 2.

Hypothesis 3 expected that ethnic majority assessors judging ethnic minority applicants would base their decision to a lesser extent on their own ratings than on ratings of others than when judging ethnic majority applicants. When addressing this hypothesis, we investigated whether assessors use their own ratings or ratings of others (i.e., from other assessors or from the applicant) when judging ethnic minority and majority applicants. More specifically, we examined whether the β weight for the interview (own rating) was smaller for ethnic minority applicants than for ethnic majority applicants, and whether the β weights for the AC- and the personality questionnaire (ratings by others) was larger for ethnic minority applicants than for ethnic majority applicants.

For the final recommendation factors Agency and Communion differences in β weights existed. For the final recommendation factors Agency as well as Communion, thirteen significant differences in β weights were found (see Tables 3 and 4). For Agency, three of the thirteen differences in β weights were on the interview. For two of these three, less weight was given to cues

when judging ethnic minority applicants. The other ten differences concerned the AC- and the personality questionnaire factors and for all of the ten occasions significantly more β weights were given to cues when judging ethnic minority applicants. For the final recommendation factor Communion, five of thirteen differences in β weights were on the interview. For four of these five, less weight was given to cues when judging ethnic minority applicants. Eight of thirteen differences on the final recommendation factor Communion concerned the AC- and the personality questionnaire factors and for all of the eight significant differences more weight were given to cues when judging ethnic minority applicants.

In sum, concerning Hypothesis 3, on the interview significantly lower weights were allocated for the judgment of ethnic minority applicants than for the judgment of majority applicants on six of eight significant differences. Contrary to the interview, on the AC- and the personality questionnaires factors significantly higher weights were allocated for the judgment of ethnic minority applicants than for the judgment of majority applicants on all of the eighteen significant differences. Thus, there seem to be systematic differences in the cue source used when coming to a recommendation for ethnic minority versus ethnic majority applicants. When judging ethnic minority applicants, selection decisions are based to a lesser extent on the interview and based to a larger extent on the AC and the personality questionnaire. Support, therefore, was found for Hypothesis 3: Ethnic majority assessors judging ethnic minority applicants base their decision to a lesser extent on their own ratings than on ratings of others than assessors judging ethnic majority applicants.

Results of the final recommendation factors Agency and Communion, concerning Hypothesis 1 to 3, firstly showed that to come to a final recommendation, the same amount or more information sources were used to judge ethnic minority applicants (Hypothesis 1). Secondly, more irrelevant cues were used for the judgment of ethnic minority applicants (Hypothesis 2). Finally, ethnic majority assessors judging ethnic minority applicants based their decision to a lesser extent on their own ratings than on ratings of others (Hypothesis 3). To come to a final recommendation on Socio-Cultural Awareness, the same (amount of) information was used when judging ethnic minority versus ethnic majority applicants. We will return to the final recommendation factor Socio-Cultural Awareness in the Discussion.

3.4 Discussion

Judgment analysis was used to investigate assessor-specific evaluations. Results largely supported the three hypotheses. Assessors used as much as or more sources of information when judging ethnic minority applicants than when judging ethnic majority applicants, which supports findings from Barr and Hitt (1986) and Singer and Bruhns (1991). Furthermore, assessors used more irrelevant information when judging ethnic minority applicants than when judging ethnic majority applicants. These results support earlier work of Ettenson et al. (1987), who pointed to the phenomenon that experienced judges appeared to be better in discriminating relevant from irrelevant information. Thus, support for Hypotheses 1 and 2 confirm the view that ethnic majority assessors can be seen as less experienced when judging ethnic minority applicants. As a result they tend to use more sources of information in their judgment of ethnic minority applicants, and also more irrelevant sources of information. Moreover, evidence was found that, when judging ethnic minority applicants, assessors gave less weight to the interview than to the AC and the personality questionnaire. As indicated before, during the selection process of the Dutch police, assessors who conduct the interview are the ones who also give the final employment recommendation. When judging ethnic minority applicants, the assessors gave less weight to their own ratings, namely the information from the employment interview. When judging ethnic minority applicants, assessors gave more weight to the AC and the personality questionnaire than to the interview, which are judgments by others (other assessors or self-ratings by applicants). It may be argued that assessors are less secure in their judgments when evaluating ethnic minority applicants. Therefore, they do not dare, as much as when judging ethnic majority applicants, to make their judgments on the basis of their own ratings from the employment interview. And, as a result, they rely more on other sources of information, such as the AC and the personality questionnaire.

Although the hypotheses were largely confirmed, the results showed differences between the three final recommendation factors. On the one hand there were the final recommendation factors Agency and Communion, which showed quite similar differences in cue weights. On the other hand, there was the final recommendation factor Socio-Cultural Awareness, which showed no differences between judgments of ethnic minority and ethnic majority applicants. An explanation for this phenomenon possibly lays in the cue that is relevant to come to this final recommendation factor Socio-Cultural Awareness. The interview appeared to be highly correlated with the Socio-Cultural Awareness factor (r = .82). Therefore, it is argued that for the

evaluation of ethnic minority as well as ethnic majority applicants it is quite clear that the interview is the main relevant cue to come to a final selection recommendation on Socio-Cultural Awareness. Because of this very obvious overlap, no differences may have existed in cue use for the evaluations of ethnic minority versus majority applicants. For the final recommendation factors Agency and Communion there are no such high correlations between the final recommendation factors on the one hand and one single cue on the other hand. Assessors, when coming to a final employment recommendation on one of the two factors, have to combine different cues and different sources of information. It seems that this process of information combining and weighting, might cause the existing differences in cue use when judging ethnic minority and majority applicants.

Although in the present study differences in cue-use between judgments of ethnic minority versus ethnic majority applicants were shown, some psychologists (seven of sixteen psychologists on the final recommendation factor Agency as well as Communion) were consistent across both ethnic groups in the use of information. For these assessors, the ethnicity of the applicant does not seem to make a difference in information processing during personnel selection. How may this be explained? A possible explanation relates to the issue of self-concept. Research on 'relative richness' of self-representation (e.g., Tversky, 1977) showed that individuals who belong to a certain group (e.g., the ethnic majority group) can be considered as experienced in judging their own group. In the domain of ethnicity, some ethnic majority group members may represent themselves particularly in terms of their ethnicity. However, it is also known that people can represent themselves in other ways (e.g., Kunda, 1999). An individual belonging to the ethnic majority group might not represent him- or herself as an ethnic majority group member, but his or her gender or age might be more important than his or her ethnicity. From this viewpoint, it may be argued that assessors who were consistent in the use of information in evaluations or ethnic majority and ethnic minority applicants, had a self-representation that was less strongly defined by ethnicity. Therefore, these assessors can just as much be seen as experienced judges of ethnic majority applicants and experienced judges of ethnic minority applicants. Further research needs to be conducted on the role of self-definition (Markus, Smith, & Moreland, 1985) and self-concept (Kunda, 1999) in selection processes.

In studies of judgment analysis, at least two points of view are seen. The first is the process view, which is concerned with how judgments are formed over time, i.e., what happens between the moment the cues are presented and the moment when a judgment is produced. What should be kept in mind is that

when using multiple regression analysis for judgment analysis, no information is given about the process during which judgments are formed. The alternative is the structural view, which focuses upon the judgmental output, the dependent variables, and tries to decompose this output in terms of the input variables, the cues. Multiple regression analysis clearly belongs to the second category. It aims at decomposing actual judgments in terms of the information available (the cues) and how this information was used in terms of weights. Our conclusion must be that for the purpose of this study, judgment analysis using multiple regression analysis provides useful accounts of human, especially individual, judgments, even though for other purposes, such as that of understanding the process of judgment, judgment analysis may be of little value.

Judgment-analysis studies focus on differences in evaluations by different assessors. With this focus, judgment-analysis research has demonstrated the existence of individual differences in the way assessors weigh and combine information about targets. Adding to this insight, the present study demonstrated differences between individual assessors in evaluations of ethnically diverse applicants. Other important aspects related to applicant ethnicity during personnel selection, such as the demographic and the perceived similarity between assessors and applicants and its effect on evaluations, were not studied. Assessors might, e.g., differ in the perception of similarity with ethnically different applicants. These differences in perceived similarity between assessors, the role of explicit and implicit attitudes towards applicants from different ethnic groups, and the effects on judgments of the assessors should get more attention in future research.

Chapter 4

Through the eyes of the assessor: Demographic and perceived similarity with regard to score differences between ethnically diverse applicants [1]

Previous research by Sacco, Scheu, Ryan, and Schmitt (2003), using multilevel analysis, found no effect of demographic (ethnic) similarity between assessor and applicant on scores given on the employment interview in the U.S. Using the same multilevel-analysis technique, the present study explored the effect of the similarity between assessor (N = 264) and applicant (N = 27,746) ethnicity on ratings given to ethnic majority and minority applicants in The Netherlands (i.e., Dutch Antilleans, Moroccans, Surinamese, and Turks) during police officer selection on the assessment center (AC), the employment interview, and the final employment recommendation. The effect was not only investigated of actual demographic similarity, i.e., ethnic similarity, between assessor and applicant but also of perceived similarity of applicants by assessors. Neither demographic nor perceived similarity was able to explain score differences between the ethnic majority and the four ethnic minority groups on the AC, the employment interview, and the final employment recommendation. Therefore, no evidence was found for (dis)similarity differentially affecting evaluations of ethnically diverse applicants during personnel selection, which confirms Sacco et al.'s previous research among U.S. ethnic groups.

[1] This chapter is submitted for publication as:
De Meijer, L. A. L., Born, M. Ph., Van Loon, H., & Van der Molen, H. T. (submitted). Through the eyes of the assessor: Demographic and perceived similarity with regard to score differences between ethnically diverse applicants.
The study in this chapter was also presented at the 22nd annual conference of the Society for Industrial and Organizational Psychology (SIOP), New York, April 2007.

4.1 Introduction

In personnel selection, differences between ethnic majority and minority groups have been widely published upon in the domain of cognitive ability (e.g., Goldstein, Zedeck, & Goldstein, 2002; Herrnstein & Murray, 1994) and, to a somewhat lesser extent, in the domain of personality (e.g., Hough, 1998; Van Leest, 1997). A characteristic that cognitive ability tests and personality questionnaires have in common, is that these can be labeled as *objective* measures, in the sense that there is no influence of a perceiving party other than the applicant him- or herself acting as a rater (cf. Bass & Barrett, 1981). In contrast, measuring devices in which a perceiving party other than the applicant him- or herself is present (e.g., an assessor, an interviewer), may be labeled as *subjective* (cf. Bass & Barrett, 1981). It is through the subjective perception by an assessor that the evaluation of an applicant takes place.

During interpersonal perception many factors may influence impressions and inferences made by a rater, among which affective processes, interpersonal factors, and motivation and skills of the rater. With regard to interpersonal factors, the similarity between the rater and the ratee may be expected to have an influence on the outcome of perceptual processes (Fiske & Taylor, 1991; Klimoski & Donahue, 2001). This similarity issue is the focus of the present study. More specifically, our study explores the relationship between ethnicity and scores on selection instruments in which judgments by ethnic majority and ethnic minority assessors are involved. The study has two main goals. The first goal is to investigate the relative extent to which demographic, in this case ethnic, similarity between assessors and applicants is able to explain existing score differences between ethnic groups on several subjective instruments. These are the assessment center (AC), the employment interview, and the final employment recommendation. The second goal is to examine the effect of perceived similarity towards ethnic groups on the scores given. Considering perceived similarity, the possible moderating role is studied of the integration into society of different ethnic groups. A large-scale dataset (N = 27,746) from the selection procedure of the Dutch police was used, containing data from ethnic majority applicants and Antillean, Moroccan, Surinamese, and Turkish minority applicants. We used multilevel analysis (MLwiN; Center for Multilevel Modeling, 1997), which is well suited for nested data structures frequently occurring in studies of demographic and perceived similarity. Both goals will now be discussed in more detail.

The first goal concerns whether *demographic similarity* – in this case actual ethnic similarity – between assessors and applicants will influence the way

assessors rate applicants. Empirical findings until now have shown mixed results concerning the effects of demographic similarity in personnel selection and on work related outcomes. Using multilevel analysis that takes into consideration the nesting of applicants within raters, Sacco, Scheu, Ryan, and Schmitt (2003) examined the demographic similarity effect on interview scores, differentiating various ethnic groups within the U.S. (i.e., White, Black, Hispanic, and Asian raters and ratees). They found no evidence that ethnic similarity played a significant role in determining the interview ratings assigned to any of the applicant groups. The present study followed the multilevel-analysis procedure used by Sacco et al. (2003) to examine the effect of both demographic and perceived similarity on scores given by assessors within a European sample. However, we extended the study of Sacco et al. (2003) to other selection measures than the interview. More specifically, we examined the AC, the employment interview, and the final employment recommendation. Sacco et al. emphasized the issues of examining different ethnic groups and perceived similarity as important directions for future research.

Returning to demographic similarity, McFarland, Ryan, Sacco, and Kriska (2004), using less sophisticated analysis-of-variance techniques, examined Black and White raters and ratees. They showed that Black raters evaluated Black applicants more favorably than White applicants, but only when the panel composition was predominantly Black. Chattopadhyay, Tluchowska, and George (2004) demonstrated in an overview of the existing literature that demographic similarity affects a range of work-related outcomes, including organizational commitment and performance.

These studies all have examined the demographic-similarity hypothesis, which states that similarity, in general, will lead to higher ratings. This expectation is derived from Social Identity Theory (SIT; Tajfel, 1982; Turner, 1987), which contends that aspects of a person's self-image come from the social categories to which he/she considers him/herself to belong to (e.g., ethnic group, gender). Social identity is seen as necessary to boost one's self-esteem. To the extent that individuals' social identities and self-categorizations are built around their demographic characteristics, demographic dissimilarity may have a negative effect on the attitudes and behaviors towards others, whereas higher identification and similarity may lead to more positive attitudes and behaviors towards other people. In line with this common idea, we also expect for our ethnic groups that demographic similarity between assessor and applicant will lead to higher ratings (Hypothesis 1).

The second goal of this study is to test the effects of *perceived intergroup*

similarity. The focus here is not on similarity in objective demographic characteristics, but on perceptions of similarity, including less tangible attributes such as values, beliefs, and personality. Most investigations of perceived-similarity effects have focused on employee relationships and performance (Ensher & Murphy, 1997; Lankau, Riordan, & Thomas, 2005; Strauss, Barrick, & Connerley, 2001; Turban, Dougherty, & Lee, 2002). As with demographic similarity, some studies (e.g., Strauss et al., 2001; Turban et al., 2002) did not use multilevel analysis, while it would have been more appropriate because of their nested data. Therefore, the question rises whether significant results were rightfully found in these studies. Nevertheless, studies justifiably using ANOVA or regression as an approach to examine data that does not show a nested structure (e.g., Ensher & Murphy, 1997; Lankau et al, 2005) have generally supported the notion that perceived similarity is positively related to relevant dependent variables (such as mentoring quality). We extend these findings to the selection context and expect that the more assessors perceive an applicant's ethnic group as similar to themselves, the higher the applicant will be rated. In the present study, we decided to go one step further and examine the differential effect of perceived similarity between ethnic minority groups. That is, we argue that perceived similarity might have a different effect for one ethnic minority group relative to another, depending on the degree to which a certain ethnic group is integrated into society. In other words, we argue that the effect of perceived similarity towards an ethnic group on evaluations given is moderated by the integration into society of that particular ethnic group. In the present study, we followed the definition of the concept integration presented by Berry (1997, p. 9): "When a person is integrated in the society in which one lives, there is an interest in maintaining one's original culture, while having daily interaction with other groups." In other words, some degree of one's own cultural integrity is maintained, while at the same time one seeks to participate in the larger societal network (Segall, Dasen, Berry, & Poortinga, 1999).

When the members of a certain ethnic minority group are isolated from the society in which they live and the general societal perception of this group is one of not being integrated, this group will be perceived as less similar relative to other – more integrated – groups. Perceptions of similarity toward a less integrated – more isolated – minority group may have a more outspoken positive effect on evaluations of applicants than perceptions of similarity toward a minority group that is more integrated and thus already is more similar to the ethnic majority group.

In The Netherlands several studies have been conducted to examine the integration hierarchy in Dutch society of different ethnic groups (e.g.,

Hagendoorn, 1995; Hraba, Hagendoorn, & Hagendoorn, 1989; Verkuyten, Hagendoorn, & Masson, 1996; Weijters & Scheepers, 2003). In this ethnic hierarchy, non-Dutch European groups were placed on top, followed by former Dutch colonial and finally Islamic groups at the bottom. The largest ethnic minority groups in The Netherlands are from the Dutch Antilles, Morocco, Surinam, and Turkey. The Dutch Antilles and Surinam are former Dutch colonies and Morocco and Turkey are (mostly) Islamic. Assuming that assessors in the selection context share the general notion about the integration hierarchy in Dutch society, it is expected that the Antillean and Surinamese groups are viewed by assessors as most integrated in Dutch society and the Turkish and Moroccan groups as least integrated (Hypothesis 2). Note that we did not include a non-Dutch European minority group. This group was too small and too diverse in our sample.

Combining what we know about perceived similarity and integration, it may be expected that the less assessors in general view an ethnic minority group as integrated, i.e., the more this group is viewed as isolated from society, the more effect an individual assessor's perceived similarity of this minority group will have on the scores given. Vice versa, it may be expected that the more an ethnic minority group is viewed as integrated into society, the less effect perceived similarity by the individual assessor of this – more integrated – minority group will have on the scores given. In sum, it is expected that the effect of perceived similarity on the scores given will be moderated by the degree of integration into Dutch society of the ethnic group (Hypothesis 3).

Finally, relating the findings on demographic and perceived similarity, several researchers found stronger effects for perceived than for demographic similarity in the domains of mentoring relationships (e.g., Ensher, Grant-Vallone, & Marelich, 2002) and performance appraisal (e.g., Strauss et al., 2001). Ferris and Judge (1991) suggest that one reason for finding stronger effects of perceptions of similarity is that people react on the bases of perceptions of reality, not on the basis of reality per se. In line with this reasoning, it is expected that perceived similarity will have a stronger effect on ratings than demographic similarity (Hypothesis 4).

4.2 Method

Participants and Procedure

Data came from 27,746 applicants who applied for a position at the Police Academy of The Netherlands from September 2001 until February 2007. Of

these, 3,089 (11%) were ethnic minority applicants. Minority applicants came from the largest ethnic minority groups in The Netherlands, namely from Dutch Antillean, Moroccan, Surinamese and Turkish ethnic groups.

Applicants who are interested in a job as police officer first apply to the local police force where they want to work after completion of their training. For the selection procedure, the local police forces routinely send all applicants to the national Police Center for Competence Assessment and Monitoring (CCM). Applicants go through two stages in the selection process. In the present study we focus on the second stage, which includes a personality questionnaire, an AC assignment and an employment interview (for an overview of the selection process, see Appendix A). The psychologist who conducts the interview is also the one who writes the final employment recommendation for the local police force. In this recommendation, the test results of the personality questionnaire, the AC ratings, and the employment interview ratings are integrated.

To investigate the effects of demographic and perceived similarity, ratings from the AC, the employment interview, and the final employment recommendation were used. In the remainder of this paper, two separate groups of raters are examined, namely the assessors who conduct the AC, and the psychologists who conduct the interview and write the final employment recommendation.

Data from 147 assessors (84% female; n = 12 belonged to the ethnic minority group) and 117 psychologists (84% female; n = 4 belonged to the ethnic minority group) were used to investigate the effect of *demographic similarity* (Hypothesis 1) on score differences between ethnic groups. In total, the assessors evaluated 26,774 applicants and the psychologists evaluated 26,588 applicants. On average, each assessor evaluated 182 applicants and each psychologist 227 applicants. Unfortunately, the number of ethnic minority psychologists was quite small. Since only a very small number of psychologists belonged to the ethnic minority group, power issues will limit the proper examination and the generalizability of the demographic-similarity effect on the ratings given by these psychologists on the employment interview and the final recommendation. On the other hand, the number of ethnic minority assessors was adequate to examine the effects of demographic similarity on ratings given by these assessors on the AC.

Related *to perceived similarity* and *integration* (Hypotheses 2 and 3), evaluations by 15 assessors (80% female; 93% ethnic majority-group member) and 12 psychologists (92% female; 100% ethnic majority-group member) were used.

In total, the assessors evaluated 6,213 applicants and the psychologists evaluated 6,879 applicants. On average, each assessor evaluated 414 applicants and each psychologist 573 applicants. With regard to perceived similarity, the group of assessors (n = 15) and the psychologists (n = 12) are sub-samples of the total group of assessors (n = 147) and psychologists (n = 117). Perceptions of similarity were available for those who filled out a perceived-similarity questionnaire. Only those assessors and psychologists were asked to fill out the questionnaire who evaluated an adequate number of ethnic majority and minority applicants (for further information regarding the perceived-similarity questionnaire, see the section 'Measures').

All raters had a high educational level (higher professional education ["HBO"] or academic-oriented education ["WO"]). Table 1 gives the sample sizes of each applicant type-rater type combination.

Table 1

Sample Sizes of Each Applicant Type – Rater Type Combination

Applicant ethnicity	Demographic similarity				Perceived similarity	
	Assessors		Psychologists		Assessors	Psychologists
	Ethnic majority	Ethnic minority	Ethnic majority	Ethnic minority	Total	Total
Ethnic majority						
Applicant *n*	20,995	2,901	22,279	543	5,390	6,128
Rater *n*	135	12	113	4	15	12
Antillean						
Applicant *n*	172	30	192	9	47	48
Rater *n*	58	7	69	3	9	11
Moroccan						
Applicant *n*	413	62	461	9	114	123
Rater *n*	85	6	86	3	13	11
Surinamese						
Applicant *n*	521	81	581	16	136	155
Rater *n*	96	8	87	4	14	12
Turkish group						
Applicant *n*	841	108	919	21	225	259
Rater *n*	105	8	97	3	15	12

Measures

Assessment Center (AC) and Employment Interview

Article 2 of the Dutch police law states: "The police force has the duty of ensuring the effective maintenance of the legal order and helping those in need." (cf. Van Loon, 2003). This definition reflects the core of the Dutch police task. Based on this article and on a thorough job analysis conducted by psychologists at the Dutch police who were experienced in job-analysis research, an AC and an employment interview had been developed to measure the following twelve dimensions: Communication Skills, Social Skills, Empathy, Initiative, Flexibility, Stress Tolerance, Emotional Stability, Authority, Decisiveness, Tolerance Towards Others, Integrity, and Self-Understanding. For an overview of these dimensions and their definitions, see Appendix B.

Assessment Center (AC). The AC is designed to evoke authentic behavior of applicants. As applicants have no relevant police work experience, the AC – more specifically, a role-play exercise in the Dutch police selection – is not a direct simulation of police work. Three people participate in each role-play, namely an actor, an assessor, and an applicant. Assessors and actors, alternately, assess and act in subsequent role-plays.

At the end of the role-play exercise, the assessor and the actor independently rate the applicant on a 7-point Likert-scale ranging from 1 (extremely weak) to 7 (excellent), on each of the following seven dimensions: Communication Skills, Social Skills, Empathy, Initiative, Stress Tolerance, Authority, and Decisiveness. After the assessor and the actor have completed their independent ratings, they discuss each applicant to reach consensus on the final dimension ratings, again, using a 7-point scale. Interrater reliabilities of the independent ratings of the assessors and the actors, given prior to the moment consensus was reached, ranged from .82 to .88 (established from a sub-sample of $N_{assessor}$ = 198 and N_{actor} = 198). Principal component analysis with varimax rotation on the consensus ratings yielded two factors, Agency and Communion (in accordance with Wiggins and Trapnell, 1996), which together explained 77% of the variance. As a measure of Agency, the average rating across the dimensions Authority, Decisiveness, Initiative, Communication Skills, and Stress Tolerance was used ($\bar{r}$ = .59; α = .87). As a measure of Communion, the average rating of the dimensions Social Skills and Empathy was used ($\bar{r}$ = .77; α = .87). In terms of behavior, Agency corresponds to the first part of Article 2 of the Dutch police law, namely: The effective maintenance of the legal order. Communion corresponds to the second part, namely: Helping those in need. The reliability of the difference

(r_{diff}) between scores on Agency and Communion is .78.

Employment Interview. The interview questions focus on evaluating behavior on the following eight dimensions: Communication Skills, Social Skills, Flexibility, Stress Tolerance, Emotional Stability, Tolerance Towards Others, Integrity, and Self-Understanding. A single psychologist conducts the interview. The interviews are semi-structured and behaviorally based, with one behaviorally anchored 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent) for each of the eight dimensions. The ratings were averaged across the eight dimensions because they were substantially correlated ($\bar{r}$ = .42; α = .85). Moreover, principal component analysis with varimax rotation yielded one interview factor that explained 50% of the variance.

Final Employment Recommendation

The final employment recommendation states to what degree an applicant is fit for a job as police officer. This recommendation is based on scores on the personality questionnaire (for a detailed description of the personality questionnaire we refer to De Meijer, Born, Terlouw, & Van der Molen [2006]), the AC, and the employment interview. After having conducted the interview with a certain applicant, the psychologist integrates the scores on the Big Five personality questionnaire, on the seven dimensions of the AC, and on the eight dimensions of the interview of this applicant into a final recommendation in terms of eleven dimensions on a 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent). These eleven dimensions are: Communication Skills (intercorrelation between AC and interview score: r = .59, p < .001), Social Skills (intercorrelation between AC and interview score: r = .40, p < .001), Empathy, Initiative, Flexibility, Stress Tolerance (intercorrelation between AC and interview score: r = .43, p < .001), Authority, Decisiveness, Tolerance Towards Others, Integrity, and Self-Understanding (for definitions, see Appendix B). The interviewer was not aware of the existing intercorrelations between the AC and the interview scores on Communication Skills, Social Skills, and Stress Tolerance (the other eight dimensions were either rated during the AC or during the interview, so for these dimensions it was not possible to calculate intercorrelations). Principal component analysis with varimax rotation on ratings on the eleven dimensions yielded three final-recommendation factors, namely Agency, Communion, and Socio-Cultural Awareness. Altogether these factors explained 67% of the variance. As a measure of Agency, the average rating across the dimensions Authority, Decisiveness, Initiative, Communication Skills, Stress Tolerance, and Flexibility was used ($\bar{r}$ = .48; α = .85). As a

measure of Communion, the dimensions Social Skills and Empathy were used ($\bar{r}$ = .66; α = .79), and for Socio-Cultural Awareness the dimensions, Tolerance Towards Others, Integrity, and Self-Understanding ($\bar{r}$ = .39; α = .65). The reliability of the difference (r_{diff}) between the scores on Agency and Communion equals .51, r_{diff} between the scores on Agency and Socio-Cultural Awareness equals .58, and r_{diff} between the scores on Communion and Socio-Cultural Awareness equals .57.

Perceived-Similarity Questionnaire

Fifteen assessors and twelve psychologists filled out a questionnaire between May and June 2004 measuring perceived similarity, which was derived from a measure by McCroskey, Richmond, and Daly (1975). In this 17-item questionnaire, assessors and psychologists filled out to what extent they perceived the average member of a particular ethnic minority group (Dutch Antilleans, Surinamese, Moroccans, and Turks) to be similar to themselves on four aspects, namely attitudes ($\alpha_{Antilleans}$ = .87, $\alpha_{Moroccans}$ = .78, $\alpha_{Surinamese}$ = .84, and α_{Turks} = .82), values ($\alpha_{Antilleans}$ = .88, $\alpha_{Moroccans}$ = .85, $\alpha_{Surinamese}$ = .90, and α_{Turks} = .87), physical appearance ($\alpha_{Antilleans}$ = .85, $\alpha_{Moroccans}$ = .79, $\alpha_{Surinamese}$ = .82, and α_{Turks} = .62), and background ($\alpha_{Antilleans}$ = .87, $\alpha_{Moroccans}$ = .77, $\alpha_{Surinamese}$ = .62, and α_{Turks} = .77). Thus, each of the seventeen items had to be filled out regarding four ethnic minority groups. A sample value-item is 'I am of the opinion that Turkish people have the same norms and values as I have' (Likert scale from 1 to 7). The same item also had to be filled out regarding the Dutch Antillean, the Moroccan, and the Surinamese group. For each ethnic minority group, the scores were averaged across the four aspects, as the intercorrelation between the similarity perceptions was quite high ($\bar{r}_{Antilleans}$ = .70, $\bar{r}_{Moroccans}$ = .64, $\bar{r}_{Surinamese}$ = .67, $\bar{r}_{Turks}$ = .67; $\alpha_{Antilleans}$ = .91, $\alpha_{Moroccans}$ = .92, $\alpha_{Surinamese}$ = .88, and α_{Turks} = .90).

Perceived Integration

An additional item on the perceived similarity questionnaire asked the assessors and psychologists to what extent they perceived the average member of a particular ethnic minority group (Dutch Antilleans, Surinamese, Moroccans, and Turks) to be integrated into Dutch society (on a 4-point scale).

Analyses

To investigate the effect of demographic similarity (Hypothesis 1) and perceived similarity (Hypothesis 3) on the scores given on the AC, the employment interview, and the final employment recommendation, hierarchical linear modeling with MLwiN 1.10 (Center for Multilevel

Modeling, 1997) was used [2]. This technique provides for a statistically accurate treatment of nested variables. Since evaluations of applicants (level 1 [L1]) involve data nested within raters (level 2 [L2]), such dependency needs to be dealt with correctly.

Hypothesis testing in MLwiN involves evaluating a series of models. We followed the procedure used by Sacco et al. (2003), which will be outlined here. We refer to level 1 (L1) or level 2 (L2) when discussing applicant and rater effects, respectively. A significant difference in deviance (-2 * log likelihood) between an initial model and a subsequent model is a prerequisite for finding significant results in this subsequent model. In the first step, which examines within- and between-group variance (equivalent to one-way ANOVA), a *null model* is tested.

$$\text{L1: } y_{ij} = \beta_{0j} + e_{ij} \quad (1)$$
$$\text{L2: } \beta_{0j} = \gamma_{00} + \mu_{0j} \quad (2)$$

The L1 equation predicts ratings received by applicants on the AC, the interview, or the final recommendation (y_{ij}) based on the mean rating (i.e., intercept) within each of the *j* raters (β_{0j}) and the error for each of *i* applicants (e_{ij}). The L2 equation models each rater's mean rating based on the grand mean (i.e., intercept; γ_{00}) and each rater's deviation (error parameter μ_{0j}). In addition, the associated variance components of the terms μ_{0j} and e_{ij} can be used to calculate the intra-class correlation (ICC), which indexes the ratio of the between-rater variance in ratings to the total variance. Barcikowski (1981) showed that even a small ICC can inflate the alpha level (type-I error) substantially. This means that even in the case of a small ICC, i.e., when raters do not differ much among each other in the ratings given, the nested data structure should be taken into account and multilevel analysis should be used.

[2] Information about the ethnicity of the applicant is explicitly given to the psychologists but not to the assessors. The reason why psychologists have knowledge about the ethnicity of the applicant during the interview and when formulating the employment recommendation, whereas assessors do not have this knowledge during the AC, is that psychologists also write the final recommendation and all information about a certain applicant is at their disposal. This is the standardized procedure during Dutch police selection. Inferences about the ethnicity of an applicant can, however, quite easily be deduced by assessors from an applicant's name and appearance.

In the second step, the first independent variable (i.e., applicant's ethnicity [x_{1ij}]) is added to the L1 equation:

$$\text{L1: } y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij} \quad (3)$$
$$\text{L2: } \beta_{0j} = \gamma_{00} + \mu_{0j} \quad (4)$$
$$\text{L2: } \beta_{1j} = \gamma_{10} + \mu_{1j} \quad (5)$$

This model is known as the *random coefficients model* because the regression coefficients β_{0j} and β_{1j} are modeled as random effects at L2 (see Equations 4 and 5). This means that, in the random coefficient model, groups of applicants (i.e., applicants rated by different assessors or psychologists) are allowed to deviate from the mean solution, not only in the intercept (γ_{00}) but also in the slope (γ_{10}). The significance of the L2 parameters (γ_{00} and γ_{10}) indicates whether ratings are significantly different from zero and whether applicant's ethnicity is related to ratings, respectively. The error parameters μ_{0j} and μ_{1j} are associated with the corresponding coefficients at L1, namely β_{0j} and β_{1j} respectively.

If the fit of the random coefficients increases significantly over and above the null model, implying that taking into account the applicant's ethnicity results in a better fit to the data, the third step involves examining whether a L2 variable (i.e., rater's ethnicity when investigating demographic similarity or rater's perceived similarity when investigating perceived similarity [x_{2j}]) predicts the variability in the intercepts of applicants' ethnicity at L1:

$$\text{L1: } y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij} \quad (6)$$
$$\text{L2: } \beta_{0j} = \gamma_{00} + \gamma_{01} x_{2j} + \mu_{0j} \quad (7)$$
$$\text{L2: } \beta_{1j} = \gamma_{10} + \mu_{1j} \quad (8)$$

This *intercepts-as-outcomes model* tests for significant differences in mean ratings as a function of rater's ethnicity or of rater's perceived similarity (γ_{01}). If the fit of the intercepts-as-outcomes model is better than the random coefficients model, the fourth and final step involves estimating the following equations:

$$\text{L1: } y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij} \quad (9)$$
$$\text{L2: } \beta_{0j} = \gamma_{00} + \gamma_{01} x_{2j} + \mu_{0j} \quad (10)$$
$$\text{L2: } \beta_{1j} = \gamma_{10} + \gamma_{11} (x_{1ij} * x_{2j}) + \mu_{1j} \quad (11)$$

This is known as the *slopes-as-outcomes model* because rater's ethnicity or rater's perceived similarity is used to predict variability in the intercepts (γ_{01}) and the slopes (γ_{11}) of applicants' ethnicity at L1. A significant γ_{11} coefficient would be evidence for a cross-level interaction, implying that ethnicity of the rater or

perceived similarity of the rater moderates the relationship between the applicant's ethnicity and the ratings given.

Concerning the integration hierarchy of the four largest ethnic minority groups in The Netherlands as viewed by assessors and psychologists (Hypothesis 2), the mean rank of each minority group and Kendall's coefficient of concordance (Kendall's *W*) were calculated. Significant differences between the mean ranks of the four groups were tested with a chi-square test.

4.3 Results

Preliminary Findings

The intra-class correlation coefficients (ICC) related to rater differences in scoring, varied between .03 and .18 (see Table 2). An ICC below .10 is viewed as a rule of thumb below which multilevel analysis is not necessary. Barcikowski (1981), nevertheless, showed that even small values of the ICC can cause a substantial increase in the chance of a type-I error to occur. Therefore, we decided to use multilevel analyses for all selection measures and both for demographic and perceived similarity, even though some ICC values were below .10.

Table 2

Intra-Class Correlations (Proportions of Variance Due to Rater Differences)

	Demographic similarity	Perceived similarity
AC		
Agency	0.08	0.07
Communion	0.06	0.03
Employment Interview	0.18	0.16
Final Recommendation		
Agency	0.12	0.14
Communion	0.09	0.05
Socio-Cult. Awareness	0.18	0.12

To get an overview of the existing score differences between the ethnic majority group and the four ethnic minority groups, the term γ_{10} (see Tables 3 and 4) is relevant because it presents whether a significant difference in scores exist. We refer to a study by De Meijer et al. (2006) for a more detailed description of the existing score differences between the ethnic groups on the selection measures used in this study and, more specifically, of the effect sizes. In this study we compared both first-generation (i.e., born outside The Netherlands) and second-generation (i.e., born in The Netherlands, but at least one of the parents is born outside The Netherlands) ethnic minority groups to the Dutch majority group. The results showed that score differences on the AC, the employment interview, and the final recommendation between the ethnic majority group and the minority groups were roughly comparable to North American research findings from the literature and varied between .02 *SD* and .68 *SD*.

Main Results

The *demographic-similarity* hypothesis stated that actual ethnic similarity between the assessor or psychologist and the applicant would lead to higher ratings (Hypothesis 1). With regard to all measures, hierarchical linear-modeling results (see Table 3) did not support Hypothesis 1.

The slopes-as-outcomes model fitted better than the intercepts-as-outcomes model only for the AC-factor Agency and only for the Surinamese group ($\Delta\chi^2$ (Δdf = 1) = 4.22). However, the estimate of the interaction term γ_{11} was not significant. Furthermore, neither concerning the AC (for the Dutch Antillean, Moroccan, and Turkish group on Agency: .00 < $\Delta\chi^2$ (Δdf = 1) < 1.22, *ns*; for Communion: .00 < $\Delta\chi^2$ (Δdf = 1) < 3.10, *ns*), nor concerning the employment interview (.00 < $\Delta\chi^2$ (Δdf = 1) < .61, *ns*), and the final employment recommendation (for Agency: .00 < $\Delta\chi^2$ (Δdf = 1) < 1.27, *ns*; for Communion: .14 < $\Delta\chi^2$ (Δdf = 1) < 2.45, *ns*; and for Socio-Cultural Awareness: .00 < $\Delta\chi^2$ (Δdf = 1) < .32, *ns*), did the slopes-as-outcomes model fit better than the intercepts-as-outcomes model. This implies that the results showed no effect of an interaction between applicants' ethnicity and assessors' or psychologists' ethnicity on the scores given.

Concerning the integration hierarchy as viewed by the group of assessors and psychologists of the four largest ethnic minority groups in The Netherlands, Hypothesis 2 stated that Dutch Antillean and Surinamese groups would be placed on top, followed by Turkish and Moroccan groups at the bottom. A chi-square test revealed that the four ethnic minority groups indeed were perceived as not being equally integrated (χ^2 (df = 3) = 36.00, p < .001).

Table 3

Demographic-Similarity Results

Applicant-group comparison		L1 parameter estimates		L2 parameter estimates	
		γ_{00} (*SE*)	γ_{10} (*SE*)	γ_{01} (*SE*)	γ_{11} (*SE*)
AC-Factor Agency					
1	Ethnic majority Dutch Antilleans	4.42** (0.08)	-0.26** (0.07)	-0.07** (0.02)	*ns*
2	Ethnic majority Moroccans	4.50** (0.05)	-0.35** (0.04)	-0.07** (0.02)	*ns*
3	Ethnic majority Surinamese	4.44** (0.05)	-0.28** (0.05)	-0.08** (0.02)	0.21 (0.12)
4	Ethnic majority Turks	4.68** (0.04)	-0.53** (0.03)	-0.08** (0.02)	*ns*
AC-Factor Communion					
1	Ethnic majority Dutch Antilleans	4.38** (0.09)	-0.19* (0.09)	0.02 (0.02)	*ns*
2	Ethnic majority Moroccans	4.56** (0.06)	-0.36** (0.06)	0.02 (0.02)	*ns*
3	Ethnic majority Surinamese	4.38** (0.06)	-0.18** (0.05)	0.01 (0.02)	*ns*
4	Ethnic majority Turks	4.61** (0.05)	-0.41** (0.04)	0.02 (0.02)	*ns*
Employment Interview					
1	Ethnic majority Dutch Antilleans	4.53** (0.06)	-0.21** (0.05)	0.06* (0.03)	*ns*
2	Ethnic majority Moroccans	4.48** (0.04)	-0.16** (0.03)	0.06* (0.03)	*ns*
3	Ethnic majority Surinamese	4.50** (0.04)	-0.18** (0.03)	0.06* (0.03)	*ns*
4	Ethnic majority Turks	4.63** (0.04)	-0.32** (0.02)	0.06* (0.03)	*ns*
Final-Recommendation Factor Agency					
1	Ethnic majority Dutch Antilleans	4.50** (0.07)	-0.32** (0.06)	0.00 (0.04)	*ns*
2	Ethnic majority Moroccans	4.51** (0.06)	-0.32** (0.04)	0.00 (0.04)	*ns*
3	Ethnic majority Surinamese	4.46** (0.05)	-0.28** (0.04)	0.01 (0.04)	*ns*
4	Ethnic majority Turks	4.71** (0.05)	-0.52** (0.03)	0.00 (0.04)	*ns*

Final-Recommendation Factor Communion					
1	Ethnic majority Dutch Antilleans	4.60** (0.08)	-0.21* (0.07)	-0.03 (0.04)	*ns*
2	Ethnic majority Moroccans	4.67** (0.06)	-0.28** (0.08)	-0.03 (0.04)	*ns*
3	Ethnic majority Surinamese	4.60** (0.06)	-0.20** (0.04)	-0.03 (0.04)	*ns*
4	Ethnic majority Turks	4.74** (0.06)	-0.34** (0.03)	-0.03 (0.04)	*ns*
Final-Recommendation Factor Socio-Cultural Awareness					
1	Ethnic majority Dutch Antilleans	4.40** (0.05)	-0.07 (0.04)	-0.03 (0.03)	*ns*
2	Ethnic majority Moroccans	4.45** (0.04)	-0.12** (0.03)	-0.03 (0.03)	*ns*
3	Ethnic majority Surinamese	4.45** (0.04)	-0.13** (0.03)	-0.02 (0.03)	*ns*
4	Ethnic majority Turks	4.49** (0.04)	-0.18** (0.02)	-0.02 (0.03)	*ns*

Note. A significant γ_{00} means that the intercept (grand mean) differs from zero. A negative γ_{10} means that ethnic minority applicants receive lower scores than majority applicants. A negative γ_{01} means that ethnic minority raters give lower scores than majority raters.
γ_{11} is the interaction of applicant and rater ethnicity, which is the focus regarding demographic similarity.
* $p < .05$ (two-tailed), ** $p < .001$ (two-tailed), *ns* means not significant (i.e., the slopes-as-outcomes model does not fit better than the intercepts-as-outcomes model).

Results in Table 5 supported Hypothesis 2 and showed the following hierarchy (with substantial agreement among the assessors and psychologists; Kendall's $W = .60$): The Surinamese minority group (mean rank = 3.60) was perceived as most integrated into Dutch society, followed by the Antilleans (mean rank = 2.80). The Turks (mean rank = 2.40) and the Moroccans (mean rank = 1.20) were perceived as the least integrated minority groups.

With regard to *perceived similarity*, it was stated that the effect of perceived similarity towards different ethnic groups on the scores given would be moderated by the degree of integration into Dutch society of ethnic groups (Hypothesis 3). Perceived similarity judgments were given for the four largest ethnic minority groups, namely Dutch Antilleans, Surinamese, Moroccans,

and Turks. With regard to the selection measures involved, hierarchical linear modeling results (see Table 4) showed no support for Hypothesis 3.

Table 4

Perceived-Similarity Results

Applicant-group comparison		L1 parameter estimates		L2 parameter estimates	
		γ_{00} (*SE*)	γ_{10} (*SE*)	γ_{01} (*SE*)	γ_{11} (*SE*)
AC-Factor Agency					
1	Ethnic majority Dutch Antilleans	4.73** (0.16)	-0.23 (0.15)	-0.08** (0.02)	*ns*
2	Ethnic majority Moroccans	4.97** (0.13)	-0.30** (0.09)	-0.15** (0.03)	*ns*
3	Ethnic majority Surinamese	4.71** (0.10)	-0.15 (0.08)	-0.10** (0.02)	*ns*
4	Ethnic majority Turks	5.10** (0.09)	-0.43** (0.07)	-0.14** (0.02)	*ns*
AC-Factor Communion					
1	Ethnic majority Dutch Antilleans	4.55** (0.18)	-0.02 (0.17)	-0.04* (0.02)	*ns*
2	Ethnic majority Moroccans	4.91** (0.15)	-0.39** (0.11)	-0.05 (0.03)	*ns*
3	Ethnic majority Surinamese	4.68** (0.12)	-0.12 (0.10)	-0.05* (0.02)	*ns*
4	Ethnic majority Turks	5.13** (0.11)	-0.50** (0.08)	-0.07** (0.02)	*ns*
Employment Interview					
1	Ethnic majority Dutch Antilleans	3.69** (0.12)	-0.10 (0.10)	0.27** (0.02)	*ns*
2	Ethnic majority Moroccans	3.87** (0.08)	-0.19* (0.06)	0.25** (0.02)	*ns*
3	Ethnic majority Surinamese	3.79** (0.08)	-0.21** (0.05)	0.26** (0.02)	*ns*
4	Ethnic majority Turks	3.97** (0.07)	-0.39** (0.04)	0.28** (0.02)	*ns*

Final-Recommendation Factor Agency					
1	Ethnic majority Dutch Antilleans	3.92** (0.15)	-0.17 (0.14)	0.16** (0.02)	*ns*
2	Ethnic majority Moroccans	4.08** (0.10)	-0.22* (0.08)	0.12** (0.02)	*ns*
3	Ethnic majority Surinamese	4.11** (0.11)	-0.36** (0.07)	0.15** (0.03)	*ns*
4	Ethnic majority Turks	4.33** (0.09)	-0.57** (0.06)	0.15** (0.03)	*ns*
Final-Recommendation Factor Communion					
1	Ethnic majority Dutch Antilleans	3.96** (0.16)	-0.04 (0.14)	0.16** (0.03)	*ns*
2	Ethnic majority Moroccans	4.25** (0.12)	-0.21* (0.09)	0.13** (0.03)	*ns*
3	Ethnic majority Surinamese	4.04** (0.13)	-0.26* (0.09)	0.20** (0.03)	*ns*
4	Ethnic majority Turks	4.23** (0.11)	-0.40** (0.07)	0.19** (0.03)	*ns*
Final-Recommendation Factor Socio-Cultural Awareness					
1	Ethnic majority Dutch Antilleans	3.74** (0.10)	-0.06 (0.08)	0.22** (0.02)	*ns*
2	Ethnic majority Moroccans	3.92** (0.07)	-0.15* (0.06)	0.19** (0.02)	*ns*
3	Ethnic majority Surinamese	3.76** (0.08)	-0.11* (0.05)	0.21** (0.02)	*ns*
4	Ethnic majority Turks	3.90** (0.07)	-0.22** (0.04)	0.22** (0.02)	*ns*

Note. A significant γ_{00} means that the intercept (grand mean) differs from zero. A negative γ_{10} means that ethnic minority applicants receive lower scores than majority applicants. A positive γ_{01} means that raters who perceive a certain ethnic minority group as more similar to themselves give higher scores than raters who perceive this ethnic minority group as less similar.
γ_{11} is the interaction of applicant ethnicity and perceived similarity toward the applicant's ethnic group, which is the focus regarding perceived similarity.
* $p < .05$ (two-tailed), ** $p < .001$ (two-tailed), *ns* means not significant (i.e., the slopes-as-outcomes model does not fit better than the intercepts-as-outcomes model).

Neither concerning the AC (for Agency: $.01 < \Delta\chi^2$ ($\Delta df = 1$) $< .88$, *ns*; for Communion: $.02 < \Delta\chi^2$ ($\Delta df = 1$) < 1.32, *ns*), nor concerning the employment

interview ($.05 < \Delta\chi^2$ ($\Delta df = 1$) < 1.18, *ns*), and the final employment recommendation (for Agency: $.10 < \Delta\chi^2$ ($\Delta df = 1$) < 1.26, *ns*; for Communion: $.08 < \Delta\chi^2$ ($\Delta df = 1$) $< .72$, *ns*; and for Socio-Cultural Awareness: $.01 < \Delta\chi^2$ ($\Delta df = 1$) < 3.03, *ns*), did the slopes-as-outcomes model fit better than the intercepts-as-outcomes model. This implies that the results showed no effect of an interaction between applicants' ethnicity and assessors' or psychologists' perceived similarity. Also, the results showed that the integration into Dutch society of the four different ethnic groups did not have a moderating role in the relationship between perceived similarity and the scores given.

Table 5

Mean Ranks in the Integration Hierarchy as Viewed by Raters

Ethnic minority group	Mean Rank
Surinamese	3.60
Dutch Antilleans	2.80
Turks	2.40
Moroccans	1.20

Note. The higher the mean rank, the more the ethnic minority group is viewed as being integrated into Dutch society.

When comparing the effects found for demographic and perceived similarity, the present study did not show stronger effects for perceived similarity than for demographic similarity. Hence, no support was found for Hypothesis 4. Both demographic and perceived similarity between assessors or psychologists and applicants showed not to have an effect on the scores given on the AC, the employment interview, and the final employment recommendation. These results were contrary to other study results in which effects for perceived similarity were found to be clearer than for demographic similarity in work-related domains such as mentoring and performance appraisal (e.g., Ensher et al., 2002; Strauss et al., 2001).

4.4 Discussion

Diversification of the workforce has become an important goal in the industrialized world. One strategy in striving for a diversified workforce

during personnel selection is to avoid the substantial adverse impact that generally is caused by cognitive ability tests (e.g., Murphy, 2002) and by using a series of face-valid non-cognitive ability selection tools without losing predictive power. Such tools include the assessment center (AC) and the employment interview. These often involve a rater who will subjectively give an evaluation of the ability, behavior, or aptitude of the applicant. Score differences between ethnic groups on these subjective measures are smaller than on the cognitive ability test. Yet, they still are quite substantial (De Meijer et al., 2006). As it is not very well known to what extent subjectivity of ratings may contribute to these score differences, we looked into effects of similarity between raters and applicants in terms of ethnicity. We investigated the effects of demographic and perceived similarity between raters and applicants on score differences on the AC, the employment interview, and the final recommendation. To this end, a distinction was made between the ethnic majority group and the four largest ethnic minority groups in The Netherlands, namely Dutch Antilleans, Moroccans, Surinamese, and Turks. Our data came from a field study in the context of personnel selection at the Dutch police ($N_{applicant}$ = 27,746). Multilevel analysis was used to deal with the nested structure of our data. One earlier study, using this same method of analysis (Sacco et al., 2003) only examined demographic similarity and yielded no effects on the scores given on the interview. The question rises whether other published research, which has analyzed demographic- as well as perceived-similarity effects at the individual level without taking into account the nested nature of the data (McFarland et al., 2004; Strauss et al., 2001; Turban et al., 2002), might have unjustly concluded that significant effects existed when there was, in fact, inadequate evidence for rejecting the null hypothesis (type-I error). This type-I error is likely to occur when analyses disregard the fact that data are structured in multiple levels, as in our study. Therefore, we believe that more credence should be given to findings from multilevel analyses.

First, our results showed no effects of demographic similarity between assessor or psychologist ethnicity and applicant ethnicity on evaluations given to ethnic majority applicants and to applicants from the four ethnic minority groups. No effects were found on the AC, the employment interview, and the final recommendation (Hypothesis 1). These results are supportive of the findings of Sacco et al. (2003).

Second, with regard to the integration of ethnic minority groups in Dutch society, we found that the Moroccan group was viewed as least integrated into Dutch society, followed by the Turkish group, the Antillean group, and finally the Surinamese group that was viewed as most integrated (Hypothesis 2).

Several studies in The Netherlands (e.g., Van Rijn, Zorlu, Bijl, & Bakker, 2004) have indeed indicated the isolated position of (mostly) Islamic minority groups, such as Moroccans and Turks. Pinto (2004) showed that Moroccans are perceived as more traditional, more religious, and more aggressive than other ethnic minority groups. These are all quite negative perceptions of the Moroccan minority group that seem to exist in Dutch society at large. Nijsten (1998) argued that the Turkish group, similarly to the Moroccan group, has a strong sense of its own Islamic culture and history. More than in the Moroccan group, however, in the Turkish group this strong sense of an own culture and history manifests itself in loyalty, cohesion, and solidarity within the group and in an avoidance of contact with other ethnic groups (Verkuyten et al., 1996).

The former colonial minority groups, i.e., the Dutch Antillean and Surinamese group, have a stronger connection with The Netherlands because of shared history and, to some extent, shared language (e.g., Hraba et al., 1989; Verkuyten et al., 1996). Because of this shared history and language, people from the Dutch Antilles and Surinam are likely to know more about The Netherlands than people from Morocco and Turkey. Hence, people from the Dutch Antilles and Surinam are also more likely to integrate easier in Dutch society.

Third, although the four ethnic minority groups were not viewed as equally integrated into Dutch society, the integration of these four minority groups did not have a moderating role in the relationship between perceived similarity and scores given by assessors and psychologists (Hypothesis 3). No effect was found of perceived similarity toward applicants on the scores given on any of the selection measures involved. It, therefore, seems that Sacco et al.'s (2003) findings of the effect of demographic similarity on interview scores in an U.S. sample are not only generalizable to other selection measures, i.e., the AC and the final employment recommendation, but also to European minority groups. Furthermore, the same results, i.e., no effects, are found for the relationship between perceived similarity and the scores given.

A first explanation for the lack of effects of demographic and perceived similarity on given scores is that, during personnel selection, raters are held accountable for their ratings and, therefore, have a strong motive to be accurate. The costs of being wrong imply that these raters will invest more effort in the judgment task than individuals in general will. A second explanation is that well-trained raters have learned to focus on a structured task and, therefore, will be less influenced by aspects of (dis)similarity. The more general question rises whether social-psychological theories on

demographic and perceived similarity are upheld when taken out of the laboratory and tested in an applied setting?

Limitations

With regard to the effect of demographic similarity, a limitation is that only a very small sample of ethnic minority psychologists (n = 4) could be included in the study. This small sample probably has suppressed any potential effects of demographic similarity between psychologists and applicants on the ratings given on the interview and the final recommendation. Fortunately, the number of ethnic minority assessors (n = 12) evaluating the AC was adequate. Although no interaction-effect of assessor ethnicity and applicant ethnicity on the scores given on the AC was found, it may be too premature to conclude that the same null-effect exists for the psychologists. Future research should try to include larger samples of ethnic minority psychologists to investigate this issue in detail.

A second limitation of our study is the lack of available predictive-validity information about the selection measures. However, all dimensions measured with the personality questionnaire, the AC, and the employment interview, as well as how they are integrated into the final recommendation form key personality characteristics and competencies for adequate police performance as indicated on O*Net (2007, May 22). Moreover, the decision to hire or reject is based directly on the final recommendation, implying important practical use of the measures.

Conclusion

In the present study, demographic and perceived similarity between ethnically diverse assessors or psychologists and ethnically diverse applicants did not affect the scores given on the AC, the employment interview, and the final employment recommendation. Therefore, no evidence was found for bias to differentially affect evaluations of ethnic majority versus ethnic minority applicants during personnel selection. These results confirm the research by Sacco et al. (2003) on demographic similarity using an U.S. sample and extend their results to the area of perceived similarity and European ethnic minority groups. Both the study by Sacco et al. (2003) and the present study emphasize the necessity to use correct methods of analysis for nested data structures.

Explanations for the lack of effects of demographic and perceived similarity on given ratings are that during personnel selection raters have a strong motive to be accurate and that well-trained raters have learned to focus on a structured task and, therefore, will be less influenced by aspects of (dis)similarity. For practitioners, these findings, fortunately, alleviate concerns

that discrimination of ethnic minority groups due to (dis)similarity may occur during personnel selection.

Chapter 5

Criterion-related validity of Dutch police-selection measures and differences between ethnic groups [1]

This study investigated the criterion-related validity of cognitive ability as well as non-cognitive ability measures and differences between ethnic majority (N = 2,365) and minority applicants (N = 682) in Dutch police officer selection. Findings confirmed the relatively low predictive validity of cognitive ability generally found for police jobs. Previous research reported no differential prediction. The present study, however, found small but systematic evidence for differences in validity for the ethnic majority and minority group of both cognitive and non-cognitive measures. For the minority group, training performance appeared to be mainly predicted by the cognitive ability test. For the majority group, cognitive ability showed very little predictive power. Non-cognitive ability variables appeared to be somewhat more predictive in this group.

[1] This chapter will be published as:
De Meijer, L. A. L., Born, M. Ph., Terlouw, G., & Van der Molen, H. T. (in press). Criterion-related validity of Dutch police-selection measures and differences between ethnic groups. *International Journal of Selection and Assessment.*
The study in this chapter was also presented at the 21st annual conference of the Society for Industrial and Organizational Psychology (SIOP), Dallas (TX), May 2006.

5.1 Introduction

In the domain of personnel selection, differences on psychological measures between ethnic majority and ethnic minority groups have been extensively investigated. Many of these studies focused on cognitive ability (or *g*), which has been found to be a consistently good predictor of job performance across a variety of occupations (Schmidt & Hunter, 1998, 2004). In particular for more complex job levels, the predictive validity of *g* is high (Hunter, 1986). Both Kanfer and Ackerman (1989), and Salas and Cannon-Bowers (2001) have shown that cognitive ability also is essential in the training context with respect to workplace learning. Other researchers have reported a strong effect of *g* in several large-scale studies in military settings on training performance (Olea & Ree, 1994; Ree, Carretta, & Teachout, 1995; Ree & Earles, 1991). At the same time, several studies (e.g., Goldstein, Zedeck, & Goldstein, 2002; Murphy, 2002; Outtz, 2002) have shown that cognitive ability tests represent the predictor that most likely will have substantial adverse impact on employment opportunities for most ethnic minority groups. Yet, evidence has been found that differences between the ethnic majority and the ethnic minority group in cognitive ability test scores are considerably larger than the differences in measures of job performance (e.g., Roth, Huffcutt, & Bobko, 2003; Waldman & Avolio, 1991).

When employers want to maximize the skill level of their employees on the one hand and diversify their workforce on the other hand, both goals cannot be achieved at the same time because of existing subgroup differences on the cognitive ability test. A possible solution for this dilemma has been sought in the use of non-cognitive ability predictors, e.g., non-cognitive dimensions measured with the assessment center (AC) and the employment interview. The AC and the employment interview are instruments that have shown smaller score differences between ethnic groups and, consequently, a lower adverse impact on employment opportunities than the cognitive ability test (De Meijer, Born, Terlouw, & Van der Molen, 2006; Murphy, 2002). This finding has been explained by the non-cognitive dimensions measured with these devices. The aim of the present study is to investigate the predictive power of cognitive and non-cognitive ability dimensions as well as their differential predictive validity in a multi-cultural setting in the context of police training at the Police Academy of The Netherlands.

Non-cognitive ability constructs may especially be useful in predicting police officer job performance. A meta-analysis of European validity studies by Salgado, Anderson, Moscoso, Bertua, De Fruyt, and Rolland (2003) showed

several remarkable findings. First, cognitive ability did not predict job performance in police occupations as well as in other occupations. Salgado et al. reported a large (corrected) predictive validity of cognitive ability for managerial occupations (r = .67; number of studies k = 6). Nevertheless, for police occupations the (corrected) predictive validity was quite low (r = .24; k = 5) and even lower than for all other occupations in the meta-analysis. Second, for training success the authors also reported the lowest predictive validity of cognitive ability for the police (r = .25; k = 3). Finally, and more in general across jobs, they showed that the predictive validity of cognitive ability was smallest for low complex jobs (r = .51) as well as for low complex training (r = .36). Other studies (Dayan, Kasten, & Fox, 2002; Hirsh, Northrop, & Schmidt, 1986; Pynes & Bernardin, 1989), not included in Salgado et al.'s meta-analysis, found that the (corrected) predictive validity of cognitive ability for law enforcement occupations was relatively low, namely between .10 and .31. Although cognitive ability is likely to be correlated with performance in virtually any job or training, in part because all jobs and trainings for these jobs call for some learning, judgment, and active information processing (Murphy, 2002), Hirsh et al. (1986) argued that non-cognitive, behavioral, dimensions, such as interpersonal skills, play a major role in the determination of police officer success. In support of this explanation, Dayan et al. (2002) reported that over 50% of the calls to police departments are about dealing with emotional situations, dealing with threatening and violent people, and settling family disputes. In addition, they found that for police performance, cognitive and non-cognitive factors had a comparable amount of predictive power.

In personnel selection, non-cognitive constructs generally are measured by means of an AC exercise, an employment interview, or a personality questionnaire. Although ACs and interviews are measurement methods that in principle can be developed to measure virtually any construct (both cognitive and non-cognitive), in the present study, the focus is on an AC and an employment interview that predominantly measure non-cognitive constructs. Therefore, they will be labeled non-cognitive measures in the remainder of this paper.

Both the AC and the employment interview have shown to have acceptable predictive validity for (police) job performance as well as (police) training success (Dayan et al., 2002; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Pynes & Bernardin, 1989; Robertson & Smith; 2001). However, there is ongoing debate about the predictive power of the personality questionnaire. One the one hand, a large meta-analysis by Barrick, Mount, and Judge (2001) showed that especially Conscientiousness is a valid predictor across jobs. On

the other hand, Murphy and Dzieweczynski (2005) and, more recently, Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007a, 2007b) argued that personality inventories almost always turn out to be fairly poor predictors of performance. They discuss three reasons why the Big Five dimensions of personality seem to have little to do with performance in most jobs. One reason is the often vague theoretical linking between personality constructs and job dimensions. Second, little is known about how to match personality constructs to jobs. Job-analysis methods have, to a large extent, focused on determining abilities and skills that are necessary for successful job performance. It is, however, not clear whether the same methods can be applied to determine which personality constructs make a difference in performing one's job. The third reason they mention for the low predictive validity of personality is that personality-related measures used in organizations have included measures of poorly defined constructs. It is likely that these three reasons apply to training performance as well, as Cortina, Doherty, Schmitt, Kaufman, and Smith (1992) found poor predictive validities of personality inventories for police training success.

In the present study, two goals are pursued. The first goal is to investigate the predictive validity of a cognitive ability test and of several non-cognitive ability selection measures (i.e., a personality questionnaire, an AC, an employment interview, and a final employment recommendation). The strength of the relationship between the cognitive ability test scores and training results will be compared to the relationship between non-cognitive ability measures and training results. The second goal is to examine potential differences in predictive validity of selection measures between the ethnic majority and the ethnic minority group.

As for the first goal, the following three hypotheses (1 a-c) are tested. First, with regard to the personality questionnaire, it is expected – in line with the results of a meta-analysis by Barrick et al. (2001) – that only the Big Five factor Conscientiousness will have a comparable predictive power to the cognitive ability test (Hypothesis 1a). It is expected that the other four Big Five factors – in line with Barrick et al. (2001), Murphy and Dzieweczynski (2005), and Cortina et al. (1992) – will show less predictive validity than the cognitive ability test (Hypothesis 1b). In correspondence with Salgado et al.'s (2003) findings on police occupations, it further is expected that the AC, the employment interview, and the final employment recommendation each will have a predictive power that is comparable to that of the cognitive ability test (Hypothesis 1c). Hypotheses 1a through 1c were examined for ethnic majority and ethnic minority trainees, separately.

As for the second goal, we investigate whether the various selection measures will show differential validity. Most research in this area has been conduced in North America (e.g., Hunter, Schmidt, & Hunter, 1979; Rotundo & Sackett, 1999) and has used cognitive ability tests as predictors. The general conclusion from this body of research has been that there is no differential validity. To our knowledge, however, little attention has been given to possible differential prediction of non-cognitive ability measures. North American studies on differential prediction typically concern cognitive test differences between native-born English-speaking ethnic minorities and Whites. While little evidence exists for test bias against U.S. ethnic minorities, Te Nijenhuis and Van der Flier (2000) argued that the U.S. differential-prediction findings cannot be directly generalized to non-native-born, non-native-language-speaking minorities in The Netherlands. For these people, who have a limited knowledge of the language and culture, as is the case for first- or even second-generation ethnic minorities in The Netherlands (Te Nijenhuis & Van der Flier, 2000) and more generally in Europe, these tests may be of limited use and therefore may show limited predictive validity. Te Nijenhuis and Van der Flier (2000) investigated the differential validity of cognitive as well as non-cognitive tests in The Netherlands. On several occasions, they indeed found evidence for differential prediction, especially with performance criteria that had lower cognitive loadings. A possible explanation was sought in the fact that these criteria were subjective evaluations containing potential criterion bias. Criterion bias implies that for ethnic minority members the focus may be on different aspects of performance than for ethnic majority members. A hypothetical example of a situation in which criterion bias could occur is when training performance of ethnic minority trainees is attributed to their decision-making skills while training performance of ethnic majority trainees is attributed to their social skills. In the present study at the Dutch police, supervisors' subjective ratings are used as training evaluations. Therefore, it is possible that criterion bias plays a role at the Dutch police as well. In correspondence with findings of Te Nijenhuis and Van der Flier (2000), it is therefore expected that differences in predictive validity between the ethnic majority and the ethnic minority group will exist both on cognitive ability and on non-cognitive ability tests (Hypothesis 2).

5.2 Method

Participants

Data came from a sample of trainees (N = 3,117; 66% male; M_{age} = 23.75, SD = 5.97), who had been admitted to the police officer training. Data were gathered from September 2001 to January 2006. The sample contains ethnic majority group members (N = 2,365; 65% male; M_{age} = 23.68, SD = 6.10), and first- as well as second-generation ethnic minority group members in The Netherlands (N = 682; 67% male; M_{age} = 24.05, SD = 5.44). First-generation ethnic minority members are born outside The Netherlands. Second-generation minority group members, in contrast to first-generation minorities, are born in The Netherlands while at least one of their parents is born outside The Netherlands. The largest ethnic minority groups in The Netherlands are the Antillean, Moroccan, Surinamese, and Turkish groups, which are equally represented in our minority sample. Of 70 trainees (2%), it was not known if they belonged to the ethnic majority or ethnic minority group. They were excluded from further analyses. The study had a longitudinal design covering about twelve months. Criteria were gathered from the police officer training about one year after the selection procedure and include evaluations of workplace performance on typical police tasks, namely: Maintaining order and helping victims.

Selection and Training at the Police Academy of The Netherlands

Applicants who are interested in a job as police officer first apply to the local police force where they want to work after they will complete their training. For the selection procedure, the local police forces routinely send all applicants to the national police Center for Competence Measurement and Monitoring (CCM). During a requirement check at the CCM, the following minimal criteria are checked on the basis of an application form: Minimal age (16 years), Dutch nationality, possession of a swimming diploma, no criminal record, and possession of a school diploma (minimal level is preparatory vocational education level B [VBO-B]). Applicants in the selection process go through two stages. During the first stage a Dutch language-proficiency test is filled in. During the second phase a physical exercise, a cognitive ability test, a personality questionnaire, an AC assignment and an employment interview are executed. The psychologist who conducts the interview is also the one who writes the final employment recommendation to the police force. For the employment recommendation, the test results of the personality questionnaire, the AC ratings, and the employment interview ratings are used. Next to the final recommendation, the final dossier to the local police forces exists of test scores of the physical exercise, the cognitive ability test, and the

language-proficiency test. On the basis of the information from the CCM, the local police force decides whether to accept or reject.

The professions for which accepted students are to be trained for are assistant police employee (two-year training), police employee (three-year training), or all-round police employee (four-year training). The trainings on these three levels are organized in the same way, i.e., three months of theoretical training is alternated with three months of on-the-job training. The theoretical knowledge gained during the first three months has to be put into practice during the later three months. Each six months are rounded off with and examination of on-the-job performance. The three trainings differ in responsibility: The more advanced a trainee is, the more responsibility (s)he will have. All trainees who finish the training will get a job as (assistant/all-round) police employee. We will now present a more detailed description of the selection measures and the criteria used.

Cognitive Ability Test

The Police Intelligence Test (PIT; Rijks Psychologische Dienst, 1975) is a cognitive ability test and consists of 107 items divided over six subtests: Verbal Comprehension, Picture Arrangement, Numerical Reasoning, Word Fluency, Spatial Ability, and Inductive Reasoning. The time limit is 51 minutes. Applicants completed the PIT in Dutch. Prior research by Lem and Van Doorn (2000) indicated alpha reliabilities varying from .69 to .87. The correlations between the subscales varied from .32 to .57. A study by Van der Maesen (1992) showed corrected predictive validity coefficients of .39 and .46 (N =162).

Personality Questionnaire

To measure the Big Five factors Extraversion, Altruism, Conscientiousness, Emotional Stability, and Intellect, the Police Personality Questionnaire (PPV; Van Leeuwen, 2000) was used. The applicants completed the PPV in Dutch. A recent progress report by Klinkenberg and Van Leeuwen (2003) indicated alpha reliabilities varying from .72 to .78. Correlations between the scales are all lower than .60. Comparison with NEO-PI-R showed observed construct validity coefficients between .17 and .58 (N = 160). A study by Lem and Van Doorn (2000) showed observed predictive validity coefficients between .15 and .43 (N = 61).

Assessment Center (AC)
A role-play exercise is utilized, in which an assessor and an actor independently make ratings on a 7-point Likert-scale ranging from 1 (extremely weak) to 7 (excellent), on each of the following seven dimensions: Communication Skills, Social Skills, Empathy, Initiative, Stress Tolerance, Authority, and Decisiveness. Interrater reliabilities ranged from .82 to .88 (N = 198). Principal component analysis with varimax rotation yielded two factors, Agency and Communion (in accordance with Wiggins and Trapnell, 1996), which together explained 77% of the variance. As a measure of Agency, the average rating across the dimensions of Authority, Decisiveness, Initiative, Communication Skills, and Stress Tolerance was used ($\bar{r}$ = .59; α = .87). As a measure of Communion, the average rating of the dimensions Social Skills and Empathy was used ($\bar{r}$ = .77; α = .87). The reliability of the difference (r_{diff}) between scores on Agency and Communion was .78.

Employment Interview
The interview questions are focused on evaluating behavior on the following eight dimensions: Communication Skills, Social Skills, Flexibility, Stress Tolerance, Emotional Stability, Tolerance Towards Others, Integrity, and Self-Understanding. A single interviewer conducts the interview. The interviews are semi-structured and behaviorally based, with one behaviorally anchored 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent) for each of the eight dimensions. The average rating across the eight dimensions was used as the dependent variable because the ratings were substantially correlated ($\bar{r}$ = .42; α = .85). Moreover, principal component analysis with varimax rotation yielded one interview factor that explained 50% of the variance.

Final Employment Recommendation
The final recommendation as to whether an applicant is fit for a job as police officer is based on results from the personality questionnaire (PPV), the AC, and the employment interview. These scores are integrated into an employment recommendation. The dimensions in the final recommendation are: Communication Skills, Social Skills, Empathy, Initiative, Flexibility, Stress Tolerance, Authority, Decisiveness, Tolerance Towards Others, Integrity, and Self-Understanding (for definitions, see Appendix B). A 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent) is used to evaluate the behavior on the eleven dimensions. Principal component analysis with varimax rotation yielded three employment-recommendation factors, Agency, Communion, and Socio-Cultural Awareness, which altogether explained 67% of the variance. As a measure of Agency, the average rating across the dimensions Authority, Decisiveness, Initiative, Communication Skills, Stress

Tolerance, and Flexibility was used ($\bar{r}$ = .48; α = .85). As a measure of Communion, the dimensions Social Skills and Empathy, were used ($\bar{r}$ = .66; α = .79) and for Socio-Cultural Awareness, the dimensions Tolerance Towards Others, Integrity, and Self-Understanding ($\bar{r}$ = .39; α = .65). The reliability of the difference (r_{diff}) between scores on Agency and Communion is .51, r_{diff} between scores on Agency and Socio-Cultural Awareness is .58, and r_{diff} between scores on Communion and Socio-Cultural Awareness is .57.

Criteria: Training Results

Supervisors were asked to rate trainees as satisfactory (1) or unsatisfactory (0) on a number of items per examination, which measured actual police work concerning 'maintaining order' (i.e., providing for public safety by maintaining order, responding to emergencies, protecting people and property, enforcing criminal laws, and identifying, pursuing, and arresting suspects and perpetrators of criminal acts [O*Net Online, 2007, January 31]) and 'helping victims' (i.e., rendering aid to accident victims and other persons requiring first aid for physical injuries [O*Net Online, 2007, January 31]). Per examination, one single supervisor observed and, subsequently, evaluated each trainee. Supervisors rated trainees' practical skills in actual police situations with actual civilians. Supervisors were trained to evaluate police trainees. All supervisors belonged to the ethnic majority group.

Each examination involved an evaluation on a number of items, among which a subset of so-called critical items. The critical items each had to be rated as being satisfactory in order to pass the examination and are descriptions of most effective behavior in a given situation. Next to the critical items, a number of remaining items as a whole had to be satisfactory scored in order to pass the examination. These focused on required daily routines. Maintaining Order (13 items) had 6 critical items on each of which the trainee should receive a satisfactory score (examples are: 'works safely', 'gives information correctly', and 'displays authority appropriately'). Of the remaining 7 items, a minimum of 5 items had to be satisfactory scored (an example is: 'checks a person's identity'). For Helping Victims (13 items), 3 items were critical (examples are: 'finds out what someone's problem is' and 'gives emotional support'). Of the remaining 10 items, a number of 7 items had to be rated as being satisfactory (an example is: 'ends the conversation properly'). If these requirements were not met, the trainee had to sit a re-examination.

Maintaining Order (13 items; α = .47 [2]) and Helping Victims (13 items; α = .58 [2]) were chosen among a series of examinations because they are two of the most important aspects of police work (cf. O*Net Online, 2007, January 31). The items of the two examinations were averaged for each separate examination. The correlation between the average scores on Maintaining Order and Helping Victims was .04 (*ns*). The 26 item-ratings were also combined into an overall training score (α = .54 [2]).

Analyses

In order to conduct correlational analysis, Structural Equation Modeling (SEM) with Amos 6.0 (Arbuckle, 2005) was used to investigate the relationships between selection measures and training criteria. Differences in correlations with regard to the ethnic majority versus ethnic minority group were tested via multi-group analyses. Furthermore, hierarchical linear regression analysis was conducted, in which scores on a certain selection measure and ethnic group membership were entered, as variables, in the first step and the interaction term in the second step. In this manner, differences between the ethnic majority and minority group in regression equations are examined. One important problem of taking ethnic group membership as part of an interaction term (group membership then becomes a moderator) into a regression equation, is that group sizes should be about the same in order to have adequate statistical power (Aguinis & Stone-Romero, 1997). In our sample, the ethnic majority group (N = 2,365) was much larger than the ethnic minority group (N = 682). Therefore, we decided to conduct the regression analyses with roughly the same group sizes. A random sample of 700 ethnic majority trainees was drawn from our original sample (SPSS 14.0, 2005), which we then compared to the 682 ethnic minority trainees.

[2] We acknowledge that the internal consistency (Cronbach's alphas) of the criteria is quite low. However, this is a common phenomenon when different items of a certain measure are behaviorally based and do not measure an underlying construct (e.g., Motowidlo, Dunnette, & Carter, 1990). The criteria used in the present study are multi-dimensional and they measure behaviors, which are related to a certain field of police work (e.g., 'maintaining order' or 'helping victims'). Test-retest estimates might be more appropriate, but they were not available.

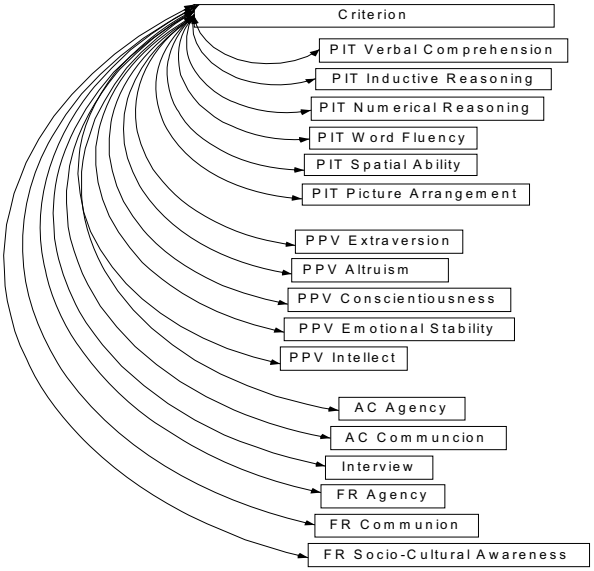
5.3 Results

Preliminary Results

Table 1 reports the alpha reliabilities, means, and standard deviations (*SD*s) of the selection measures and the criteria and the correlations among the selection measures for the ethnic majority and minority groups, separately. All selection measures had good reliabilities. Therefore, the correlations between the selection measures were not corrected for attenuation (the correlations between the selection measures and the training criteria and among the training criteria were).

Main Results

In investigating the predictive validities of all selection measures, the predictive power of the non-cognitive ability measures were compared to the predictive power of the cognitive ability test (Hypotheses 1 a-c). Simultaneously, it was examined whether differences existed between predictive validities for the ethnic majority versus ethnic minority group (Hypothesis 2). Correlations between predictor scores and criterion scores were generated by means of SEM (see Figure 1). SEM enabled the investigation of differential prediction by means of multi-group analysis. The fit indices of the models for the three criteria are shown in Table 2. The models showed a good fit.



Note. Criteria are the overall training score, Maintaining Order, and Helping Victims. PIT is the cognitive ability test, PPV is the personality questionnaire, AC is the assessment center, and FR is the final recommendation.

Figure 1. Model for correlational analysis

Table 1

Descriptives of Selection Measures and Training Criteria

Selection Measures and Criteria	Cronbach's α		Mean (*SD*)		Correlation Coefficients					
	Ethnic majority	Ethnic minority	Ethnic majority	Ethnic minority	1	2	3	4	5	6
Cognitive ability measure:										
PIT										
1. Verbal Comprehension	0.62	0.68	10.25 (2.25)	9.53 (2.45)	-	0.49**	0.45**	0.48**	0.39**	0.41**
2. Inductive Reasoning	0.70	0.75	11.03 (2.30)	10.41 (2.45)	0.43**	-	0.59**	0.37**	0.58**	0.49**
3. Numerical Reasoning	0.82	0.80	8.29 (3.28)	7.39 (3.33)	0.43**	0.52**	-	0.39**	0.46**	0.38**
4. Word Fluency	0.81	0.79	9.74 (3.54)	8.75 (3.44)	0.40**	0.34**	0.41**	-	0.34**	0.37**
5. Spatial Ability	0.81	0.83	20.87 (4.08)	19.68 (4.22)	0.34**	0.53**	0.38**	0.28**	-	0.39**
6. Picture Arrangement	0.64	0.75	8.32 (1.60)	8.05 (1.74)	0.28**	0.36**	0.25**	0.26**	0.29**	-
Non-cognitive ability measures:										
PPV										
7. Extraversion	0.76	0.71	6.30 (2.67)	6.30 (2.76)	0.02	-0.01	-0.02	0.01	-0.01	0.01
8. Altruism	0.75	0.75	5.84 (2.74)	5.98 (2.85)	-0.03	-0.02	-0.05*	-0.03	0.01	-0.01
9. Conscientiousness	0.70	0.71	5.56 (2.91)	5.90 (2.85)	-0.09**	-0.07*	-0.07*	-0.08*	-0.02	-0.06*
10. Emotional Stability	0.74	0.74	5.91 (2.74)	6.09 (2.82)	-0.03	0.00	-0.01	0.02	0.02	0.02
11. Intellect	0.77	0.75	6.07 (2.75)	5.96 (2.89)	0.08*	0.05*	0.01	0.07*	0.06*	0.06*
AC										
12. Agency	0.87	0.86	4.69 (0.83)	4.51 (0.84)	0.13**	0.10**	0.08*	0.12**	0.05*	0.14**
13. Communion	0.87	0.87	4.52 (1.03)	4.36 (1.06)	0.09**	0.08*	0.04*	0.09**	0.05*	0.10**
14. Employment Interview	0.85	0.85	4.65 (0.52)	4.56 (0.55)	0.13**	0.08*	0.05*	0.09**	0.03	0.12**
Final Recommendation										
15. Agency	0.85	0.84	4.66 (0.67)	4.49 (0.69)	0.14**	0.11**	0.08*	0.12**	0.05*	0.14**
16. Communion	0.78	0.79	4.60 (0.83)	4.48 (0.82)	0.09**	0.06*	0.03	0.07*	0.03	0.10**
17. Soc.-Cult. Awareness	0.64	0.67	4.52 (0.51)	4.48 (0.56)	0.11**	0.05*	0.02	0.07*	0.01	0.10**

Training Criteria:										
18. Overall training score	0.54	0.53	0.91 (0.13)	0.91 (0.12)	*0.04*	*0.04*	0.05	*0.06*	0.00	0.11*
19. Maintaining Order	0.47	0.53	0.91 (0.14)	0.91 (0.15)	*0.02*	*0.02*	0.02	*0.00*	-0.04	*0.08*
20. Helping Victims	0.58	0.52	0.90 (0.15)	0.91 (0.14)	*-0.04*	0.03	0.06	*0.06*	0.00	*0.07*

Table 1 – continued

	Correlation Coefficients													
	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1.	0.14**	0.04	-0.13*	0.02	0.20**	0.26**	0.17**	0.23**	0.26**	0.17**	0.18**	*0.28***	*0.20**	*0.24**
2.	0.11*	0.06	-0.06	0.09*	0.16**	0.27**	0.18**	0.18**	0.26**	0.16**	0.11*	*0.21**	*0.22**	0.10
3.	0.06	0.04	-0.05	0.08*	0.11*	0.18**	0.12*	0.12*	0.17**	0.12*	0.09*	0.12*	0.11	0.09
4.	0.17**	0.05	-0.11*	0.11*	0.22**	0.18**	0.09*	0.15**	0.18**	0.09*	0.10*	*0.23***	*0.18**	*0.20***
5.	0.11*	0.10*	0.03	0.09*	0.17**	0.13*	0.10*	0.06	0.11*	0.08*	0.03	0.05	0.04	0.05
6.	0.09*	0.05	-0.09*	0.07	0.19**	0.20**	0.13*	0.19**	0.21**	0.13*	0.16**	0.06	*-0.05*	*0.19**
7.	-	0.58**	0.33**	0.38**	0.53**	0.17**	0.13*	0.24**	0.22*	0.19**	0.12*	0.05	0.12	-0.06
8.	0.56**	-	0.44**	0.35**	0.54**	0.05	0.06	0.13*	0.08*	0.13*	0.12*	0.05	0.13	0.03
9.	0.26**	0.37**	-	0.29**	0.40**	-0.08*	-0.04	-0.07	-0.08*	-0.04	-0.06	0.08	0.10	0.00
10.	0.34**	0.34**	0.29**	-	0.40**	0.07	0.05	0.13*	0.11*	0.09*	0.01	-0.02	0.04	-0.12
11.	0.49**	0.51**	0.42**	0.41**	-	0.21**	0.11*	0.24**	0.27**	0.14**	0.13*	-0.05	0.11	*-0.13*
12.	0.12**	0.03	0.01	0.08*	0.17**	-	0.62**	0.58**	0.91**	0.59**	0.42**	*-0.01*	-0.08	*0.07*
13.	0.09**	0.04*	0.00	0.05*	0.09**	0.63**	-	0.41**	0.56**	0.89**	0.33**	*0.02*	*-0.08*	0.08
14.	0.24**	0.14**	0.04*	0.18**	0.24**	0.50**	0.38**	-	0.75**	0.56**	0.82**	*0.02*	*-0.06*	0.09

15.	0.20**	0.08*	0.04*	0.16**	0.25**	0.90**	0.56**	0.72**	-	0.60**	0.52**	*0.02*	-0.06	*0.09*
16.	0.18**	0.11**	0.02	0.08*	0.14**	0.60**	0.89**	0.56**	0.62**	-	0.47**	*-0.04*	*-0.15*	0.09
17.	0.13**	0.12**	0.03	0.08*	0.16**	0.36**	0.32**	0.80**	0.50**	0.48**	-	*0.01*	-0.05	*0.02*
18.	0.05	0.03	0.03	-0.08	0.03	*0.13***	*0.11**	*0.16***	*0.17***	*0.17***	*0.17***	-	1.00**	1.00**
19.	0.09*	0.10*	0.03	-0.04	0.05	-0.02	*0.06*	*0.04*	0.02	*0.06*	0.02	1.00**	-	0.16*
20.	0.02	-0.03	0.05	-0.04	*0.03*	*0.21***	0.12*	0.17**	*0.21***	0.18**	*0.19***	1.00**	0.05	-

Note. PIT scores varied between 1 and 15 on Verbal Comprehension and Inductive Reasoning, between 1 and 13 on Numerical Reasoning, between 1 and 25 on Word Fluency, between 1 and 29 on Spatial Ability, and between 1 and 10 on Picture Arrangement; PPV scores varied between 1 and 10; AC, interview, and final-recommendation scores varied between 1 and 7; and training scores varied between 0 and 1.

Positive *d* values indicate the ethnic majority group scoring higher.

The correlations of the ethnic minority group are above the diagonal. Correlations between selection measures are corrected for range restriction. Correlations between selection measures and criteria are corrected for range restriction and attenuation. Significant differences between ethnic groups of correlations between selection measures and criteria are in *italic* ($p < .05$). Correlations between training criteria are corrected for attenuation.

* $p < .05$, ** $p < .001$.

Table 2

Correlational Analysis: Fit Indices for Training Criteria

Criteria	χ^2	*df*	χ^2/df	TLI	CFI	RMSEA
Overall training score	764.32**	192	3.98	0.96	0.98	0.03
Maintaining Order	764.51**	192	3.98	0.96	0.98	0.03
Helping Victims	764.22**	192	3.98	0.96	0.98	0.03

Note. TLI means Tucker-Lewis Index, CFI means Comparative Fit Index, RMSEA means Root Mean Square Error of Approximation.
** $p < .001$ (one-tailed).

Table 1 shows SEM results and presents the correlations between predictors and criteria. Only the correlations corrected for direct range restriction and attenuation (for the formulae see Bobko, Roth, & Bobko, 2001) are shown (for the uncorrected correlations, the first author may be contacted). Furthermore, significant correlation differences between the ethnic majority and minority group are marked. For reasons of clarity and conciseness, an overview of the most remarkable results will be described here. First, a comparison in predictive validity will be made between the cognitive ability test and the personality questionnaire. Second, the predictive validity of the cognitive ability test will be compared to the predictive validity of the AC, the employment interview, and the final recommendation, separately.

Regarding the personality questionnaire, we expected that Conscientiousness would have a predictive power comparable to the cognitive ability test (Hypothesis 1a). Furthermore, the other Big Five factors, namely Extraversion, Altruism, Emotional Stability, and Intellect, were expected to show less predictive power than the cognitive ability test (Hypothesis 1b). Hypothesis 2 predicted that differences in predictive validities between the ethnic majority and the ethnic minority group would exist on the cognitive ability test and on the personality questionnaire. The results in Table 1 show support for Hypothesis 1b, but not for Hypothesis 1a. No support was found for Hypothesis 2 on the personality questionnaire, but support for Hypothesis 2 was found on the cognitive ability test.

All five personality factors showed very little predictive validity. Conducting multi-group analyses, a significant difference in predictive validity between the ethnic majority and minority group was found only for Intellect predicting the

training score of Helping Victims ($r_{maj.}$ = .03, *ns* and $r_{min.}$ = -.13, *ns*, respectively). No evidence for differential prediction was found for the other Big Five factors on any of the criteria.

The predictive validity of the cognitive ability test was higher than the above-mentioned predictive validity of the personality questionnaire, especially for the ethnic minority group. More specifically, the verbal subtests of the cognitive ability test (i.e., Verbal Comprehension and Word Fluency) were most predictive of training success for the ethnic minority group compared to the ethnic majority group. Significant differences in validity between the ethnic groups were found for several cognitive ability subtests, namely Verbal Comprehension, Inductive Reasoning, and Word Fluency for the prediction of the overall training score. For training results on Maintaining Order, differences in validity were found for the sub-tests Verbal Comprehension, Inductive Reasoning, Word Fluency, and Picture Arrangement. Finally, for the training scores on Helping Victims, different validity coefficients were found for the sub-tests Verbal Comprehension, Word Fluency, and Picture Arrangement. No differences in prediction were found for the sub-tests Numerical Reasoning and Spatial Ability.

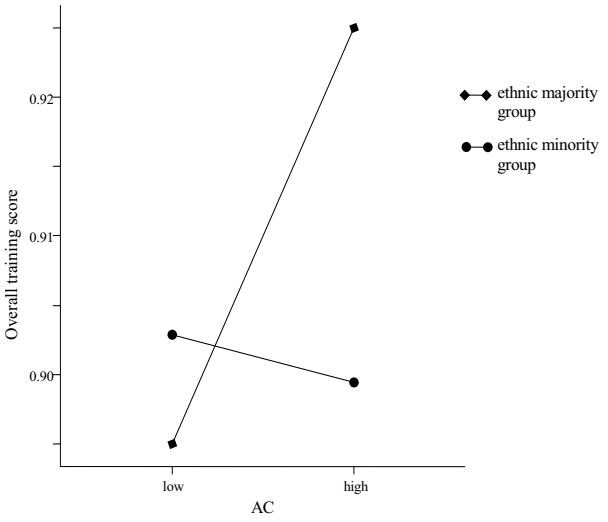
Hypothesis 1c predicted that the AC, the employment interview, and the final employment recommendation would have a predictive power comparable to the cognitive ability test. Hypothesis 2 predicted that differences in predictive validity between the ethnic majority and the ethnic minority group would exist on the AC, the interview, and the employment recommendation. The results in Table 1 show partial support for Hypothesis 1c and support for Hypothesis 2. The predictive power of the AC, the interview, and the final employment recommendation was larger than the predictive power of the cognitive ability test, but only for the ethnic majority group. For the ethnic minority group, the non-cognitive ability tests showed very small predictive validities for the overall training score as well as for training scores on Maintaining Order and Helping Victims.

Regarding the differences in predictive-validity coefficients between the ethnic groups (Hypothesis 2), the AC, the interview, and the final recommendation showed larger predictive validities for the ethnic majority group than for the ethnic minority group. The results in Table 1 showed differential validity for the overall training score of both AC-factors, the interview, and all final-recommendation factors. For training results on Maintaining Order, differences in validity coefficients were found of the AC-factor Communion, the interview, and the final-recommendation factor Communion. The other AC and final-recommendation factors showed no

differential validity for Maintaining Order. For the training results on Helping Victims, again, several selection factors showed differences in validity. Differential validity was found of the AC-factor Agency and on the final-recommendation factors Agency and Socio-Cultural Awareness. The interview and the AC- and final-recommendation factor Communion showed no differences for Helping Victims.

Concerning Hypothesis 2, hierarchical regression analysis was conducted in addition to correlational analyses. Hierarchical regression is an often-used technique to examine differential validity. Scores on a certain selection measure and group membership were entered, as variables, in the first step of the regression. The interaction between both was entered in the second step. A significant interaction effects shows evidence for differential validity. The results are shown in Table 3 (results on sub-test or sub-dimension level are not shown in Table 3, but are only described in the text). Although the incrementally explained variances of the interaction terms are very small or close to zero, significant interaction effects were found for the cognitive ability test, the AC, the employment interview, and the final recommendation. These results, thus, point to the existence of differential validity of both the cognitive ability test and the non-cognitive ability measures (Hypothesis 2).

The regression of the overall training score on the AC (for an illustration, see Figure 2), the employment interview, and the final recommendation differed for the two ethnic groups, with an only marginal difference for the employment interview.



Note. In this example, the significant interaction effect of AC scores and ethnic group membership on the overall training score is illustrated. Training scores varied between 0 and 1.

Figure 2. Illustration of ethnic majority and minority AC scores predicting the overall training score

Although the regression of the overall training score on the cognitive ability test in general did not show differences, the regression on the sub-tests Verbal Comprehension (β = -.25, p = .10 [marginally]) and Word Fluency (β = -.27, p < .05) did. The regression of Maintaining Order in general did not show differences between the ethnic majority and minority group. However, the regression on the cognitive ability sub-test Word Fluency (β = -.35, p < .05) did, as well as on the AC- and final-recommendation factor Communion (β = .29, p < .10 [marginally] and β = .43, p < .05, respectively). The regression of Helping Victims on the cognitive ability test (marginally), the AC, and the final recommendation (marginally) differed for the two ethnic groups. Especially, the regression on the cognitive ability sub-test Verbal Comprehension (β = -.34, p < .05) appeared to be different for the two groups.

Although the effect sizes of differential validity generally are small, the following trend is discernible: The cognitive ability test, especially the verbal sub-tests, appears to show more predictive power for the ethnic minority group than for the ethnic majority group. Contrarily, the AC, the employment interview, and the final employment recommendation appear to show more predictive power for the ethnic majority group than for the ethnic minority group. The personality questionnaire showed very little predictive power for either group.

Table 3

Hierarchical Regression Analyses of the Selection Measures Predicting the Training Scores

	Selection Measures									
	Cognitive ability test		Personality questionnaire		AC		Employment interview		Final recommendation	
Criteria	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
Overall training score:										
Selection measure	0.07*	0.19*	0.00	0.02	0.08*	-0.11	0.08*	-0.06	0.09*	-0.12
Ethnic group membership	-0.01	0.23	0.01	0.03	0.00	-0.34*	0.00	-0.38	0.00	-0.54*
Interaction		-0.28		-0.03		0.41*		0.42†		0.60*
ΔR^2		0.00		0.00		0.01*		0.00†		0.00*
Maintaining Order:										
Selection measure	0.05	0.15†	0.04	0.13	0.01	-0.11	0.01	-0.04	0.01	-0.11
Ethnic group membership	0.02	0.23	0.03	0.12	0.03	-0.20	0.03	-0.13	0.03	-0.31
Interaction		-0.25		-0.13		0.28		0.18		0.37
ΔR^2		0.00		0.00		0.00		0.00		0.00
Helping Victims:										
Selection measure	0.06*	0.23*	-0.03	-0.09	0.12**	-0.06	0.10*	0.00	0.11**	-0.05
Ethnic group membership	-0.03	0.30†	-0.02	-0.08	-0.04	-0.37*	-0.03	-0.32	-0.03	-0.48†
Interaction		-0.40†		0.09		0.40*		0.32		0.49†
ΔR^2		0.00†		0.00		0.00*		0.00		0.00†

Note. For these analyses, roughly equal group sizes were used (i.e., N = 700 for the ethnic majority group; N = 682 for the ethnic minority group). Standardized regression weights are presented. The crosses and asterisks correspond to the unstandardized regression weights. Ethnic group membership is coded as follows: 1 = 'ethnic minority group'; 2 = 'ethnic majority group'.
† $p < .10$, * $p < .05$, ** $p < .001$.

5.4 Discussion

As a first goal, the criterion-related validity of both cognitive and non-cognitive ability selection measures was investigated for training performance of police trainees. Second, differential prediction between ethnic groups of both cognitive and non-cognitive ability measures was examined.

When score differences between ethnic groups on a cognitive ability test are larger than score differences in job or training performance, potentially good employees or trainees could be rejected during selection. A potential problem is a lack of ethnic diversity or heterogeneity in one's workforce. Especially the latter issue is of concern for organizations in the public domain such as the police, since contact with different ethnic groups in society forms an important aspect of the job as police officer. When non-cognitive ability measures are available that show less score differences between ethnic groups than on the cognitive ability test, as in the present study, and that show comparable predictive validities to the cognitive ability test, this could be a solution. The predictive validities of the non-cognitive ability measures and the differential prediction that were found will be discussed below.

Confirming the viewpoint of Murphy and Dzieweczynski (2005) and Morgeson et al. (2007a, 2007b), the Big Five personality questionnaire showed almost no predictive power. Cortina et al. (1992) found similar results. They used a sample of police recruits and found poor predictive validities of two personality inventories for police training performance. Cortina et al. argued that the questionnaires they used were not useful for the police selection, because the tests were not developed specifically for the police. Although the personality questionnaire in the present study was adapted for the Dutch police, it is recommended to further investigate whether the constructs that are measured with the Police Personality Questionnaire (PPV) indeed are important for police training performance. In line with suggestions by Hattrup, Rock, and Scalia (1997), it can also be argued that the Big Five personality constructs do not predict the scores on the specific police-relevant criteria that were used in the present study. They may, however, be useful to predict more general training performance, e.g., teamwork, friendliness, and punctuality. Unfortunately, these criteria could not be directly investigated in the present study.

The corrected predictive validity of cognitive ability for training performance as reported in the literature by Schmidt and Hunter (1998, 2004) is high (r = .59). However, Salgado et al. (2003) reported much lower (corrected)

predictive power for cognitive ability tests in police occupations, namely .24. The present study found even lower (corrected) predictive validities of cognitive ability for training performance than Salgado et al. (2003) did, namely .04 ($-.04 < r_{maj.} < .11$) for the ethnic majority group and .14 ($-.05 < r_{min.} < .28$) for the ethnic minority group. One possible explanation for the relatively low validities of cognitive ability tests lies in the potential role of non-cognitive factors in the determination of performance in police work as stated by, e.g., Hirsh et al. (1986). Interestingly, however, in the present study this explanation will pertain more to the ethnic majority group than to the ethnic minority group. For the majority group various factors measured during the AC, the interview, and the final employment recommendation, i.e., Agency, Communion, and Socio-Cultural Awareness, were more predictive than cognitive ability for several training criteria. Especially the Agency factor of the AC and the final recommendation appeared to be predictive for the ethnic majority group. For the minority group, the cognitive ability test was most predictive, especially the verbal cognitive ability subtests, i.e., Verbal Comprehension and Word Fluency. The non-cognitive ability tests showed very little predictive power for the minority group.

The results of the hierarchical regression analyses also point to differential validity for all selection measures except for the personality questionnaire. Training performance appeared to be somewhat better predicted by several cognitive ability subtests for the ethnic minority group, and somewhat better predicted by the non-cognitive ability tests for the ethnic majority trainees. Where differences in predictive validity were found, these might have been caused by ethnic bias of ethnic majority supervisors' subjective evaluations (Te Nijenhuis & Van der Flier, 2000), even though evaluations of trainees during the Dutch police training were structured according to evaluation forms. For ethnic majority trainees, relatively more attention may have been given to the non-cognitive ability aspects of performance, i.e., social skills, decisiveness, and authority, measured with the AC, the interview, and the final-recommendation. While for ethnic minority trainees, relatively more attention may have been given to the verbal cognitive ability aspects of performance. The question remains whether supervisors' evaluations of ethnic minority trainees are predominantly susceptible to these quite basic language skills to the extent that these skills will overshadow other important non-cognitive factors, such as social skills and decisiveness. To better understand potential supervisors' susceptibility to ethnic bias, research using ethnic majority as well as ethnic minority supervisors should get more attention in the future.

Limitations

Although the total sample of ethnic minority trainees was very acceptable (N = 682), a first limitation of the present study was that this sample was too small to differentiate among ethnic minority groups. Treating ethnic minorities as a homogeneous group that merely contrasts with the ethnic majority group ignores the many visible and cultural differences among ethnic minority groups that may affect score differences, predictive validity coefficients, and differential prediction. In the present study, we extended previous studies by examining the predictive validity of a cognitive ability test as well as several non-cognitive selection measures in a multi-cultural setting. Since we found differential prediction on all measures, future research should investigate this differential prediction for the various ethnic groups that exist in The Netherlands and, more broadly, in multicultural societies, also for other sets of tests.

Second, although correlations between the cognitive ability test on the one hand and the AC, the employment interview, and the final recommendation on the other hand were all below .14 ($\bar{r}$ = .08) for the ethnic majority group and were all below .27 ($\bar{r}$ = .15) for the ethnic minority group, there appears to be some overlap between cognitive ability and the non-cognitive constructs. This might slightly contaminate the predictive validities found in the present study. In general, it is to be expected that scores on a cognitive ability test are correlated with AC and interview scores, in general, because performance on an AC or an interview to some extent requires cognitive skills such as active information processing and adequate responding (cf. Murphy, 2002).

A third limitation of the present study was that the sizes of the predictive validities were quite small. An explanation for this finding may be found in the low variance in training scores (see criteria-*SD*s in Table 1). As a result of this low variance, the correlations and regression weights presented in this study may be somewhat underestimated. On the one hand, low criterion variance may be a valid explanation for possibly underestimated predictive validities. On the other hand, previous research has also found relatively small (corrected) predictive validities of the cognitive ability test and the personality questionnaire for low-level police training and work performance (Cortina et al., 1992; Salgado et al., 2003). As the issue of relatively low criterion variance is a general problem encountered in operational criteria (Cascio & Aguinis, 2005), we believe that the predictive validities of the cognitive and non-cognitive selection measures for low-level police training as found in the present study are not less accurate estimates than other reported findings in the literature. Moreover, in our opinion, the relatively small validities are

informative, firstly, because they are systematic. The predictive validities to out view also are informative, since the goals of the present study were aimed at investigating the *differences* in predictive power of various selection measures as well as the *differential* prediction of these measures for different ethnic groups.

A final limitation is the use of training performance as a criterion instead of job performance. The question rises whether training performance can be generalized to work performance, as predicting work performance is the ultimate goal of personnel selection. Using training performance as a performance criterion may be deficient because the goal of personnel selection is to select potentially good employees, not necessarily good trainees. In a study conducted by Salgado et al. (2003), however, the validity of cognitive ability when predicting police training performance was almost equal to the validity when predicting police job performance. The issue of generalizability then concerns the non-cognitive predictors and the potential difference in predicting training performance versus job performance. As the present study has used an on-the-job performance criterion, it is expected that the validity coefficients for the training-performance criterion can largely be generalized to job performance.

Conclusion

The predictive validities of the various selection measures are roughly in line with previous research. Regarding differential prediction between ethnic groups we found somewhat different results compared to previous, mostly North American, studies. The effect sizes with regard to differential validity are small but systematic. The cognitive ability test appeared to show somewhat larger predictive-validity coefficients for the ethnic minority group than for the ethnic majority group. The non-cognitive ability measures appeared to show somewhat larger predictive validities for the ethnic majority group than for the minority group. These results may imply that it is important to use both cognitive ability and non-cognitive ability tests for the selection of police officers in order to obtain a diverse ethnic work environment.

Chapter 6

The construct-driven development of a video-based situational judgment test for integrity: A study in a multi-ethnical setting [1]

In a field study conducted in a multi-ethnical selection setting at the Dutch police, we examined the construct validity of a video-based situational judgment test (SJT) aimed to measure the construct of integrity. Contrary to previous viewpoints, we argue that a construct-driven approach may be fruitful in the development of SJTs to measure one single construct. Confirming our expectations, we found support for the construct validity of the Integrity-SJT, including only a very small relationship between scores on this SJT and cognitive ability. These results held across ethnic majority and ethnic minority applicants. Furthermore, we investigated the SJT score difference between the ethnic majority and the ethnic minority group. The results showed a SJT score difference of .38 SD. This difference is smaller than the score difference that is generally found on cognitive ability tests, which are often used in high-stakes testing. The Integrity-SJT, therefore, is a promising test for personnel selection in a multi-ethnical setting.

[1] This chapter is submitted for publication as:
De Meijer, L. A. L., Born, M. Ph., Van Zielst, J., & Van der Molen, H. T. (submitted). The construct-driven development of a video-based situational judgment test for integrity: A study in a multi-ethnical setting.

6.1 Introduction

Although situational judgment tests (SJTs) have been in use since the 1920s, they have become increasingly popular in personnel selection and in the research literature during the last two decades (e.g., Chan & Schmitt, 1997, 2005; Dalessio, 1994; McDaniel, Hartman, Whetzel, & Grubb, 2007; Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, & Donovan, 1998; Weekley & Jones, 1997, 1999). Several characteristics of the SJT have caused its revival. First, McDaniel et al. (2007) meta-analytically showed the criterion-related validity and the incremental validity of SJTs over and above a composite of cognitive ability tests and personality questionnaires in predicting job performance. Second, SJTs have been found to have less adverse impact against ethnic minority groups than more traditionally used cognitive ability tests (Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001; Motowidlo, Dunnette, & Carter, 1990; Nguyen & McDaniel, 2003; O'Connell, Harman, McDaniel, Grubb, & Lawrence, 2007; Olson-Buchanan et al., 1998; Weekley & Jones, 1997, 1999). Finally, new technology has made the development of SJTs based on video material possible. The video-based SJT appears to have several advantages compared to the paper-and-pencil SJT, such as a higher criterion-related validity (Lievens & Sackett, 2006), less adverse impact, and higher realism of the test leading to more reliable respondent reactions (Chan & Schmitt, 1997; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000).

Even though SJTs have a series of advantages, important questions still persist. A critical issue is the often-found construct-heterogeneity of SJTs and the difficulty of developing a SJT that measures one specific construct. In the research literature, a substantial debate exists concerning what SJTs actually measure. Broadly, two movements can be distinguished. On the one hand, there is the viewpoint that there is a single situational judgment construct, i.e., job knowledge (Schmidt & Hunter, 1993). On the other hand, there is a group of researchers who think that this reasoning is misguided and argue that SJTs are merely measurement methods (Chan & Schmitt, 1997; McDaniel, Morgeson, Bruhn Finnegan, Campion, & Braverman, 2001; McDaniel & Nguyen, 2001; McDaniel & Whetzel, 2005; Weekley & Jones, 1999). This latter view implies that like other measurement methods, such as the employment interview, SJTs can be built to measure a variety of constructs. For example, to assess conscientiousness, one could build a SJT where conscientiousness is the major determinant of individual differences in item responding. These researchers argue that there are, however, limits to what constructs a SJT can or cannot measure. Summarizing the empirical literature, McDaniel and Nguyen (2001) showed that SJTs are not

unidimensional construct tests, but should be considered as a measurement method capable of measuring several constructs. According to these authors, empirical evidence indicates that the constructs measured by SJTs can be limited to cognitive ability or *g,* conscientiousness, agreeableness, and emotional stability. Although we agree with the viewpoint that SJTs are measurement methods, we question whether SJTs are limited to measuring *g,* conscientiousness, agreeableness, or emotional stability. We argue that SJTs can also be built to measure other constructs. To this end, we will first describe the characteristics of SJTs in general, followed by a discussion of the concept of integrity and of integrity tests. The concept of integrity is central to our study. Finally, a SJT aimed to measure integrity will be introduced, followed by an overview of our hypotheses.

Characteristics of SJTs

SJTs typically consist of hypothetical scenarios describing a work situation in which a problem has arisen. The work situation may be a possible actual situation on the target job or a situation constructed in such a manner that it is psychologically identical to an actual work situation (Chan & Schmitt, 1997). Work situations within the test are usually developed on the basis of a critical-incident analysis involving subject matter experts (SMEs). In case of developing a SJT aimed to measure one specific construct, the SMEs are asked for critical incidents in terms of this specific construct, instead of the more general work context.

Although all SJTs have similarities, such as the fact that they consist of hypothetical scenarios as was described above, they can vary in terms of format, namely from paper-and-pencil tests with written descriptions of situations (Chan & Schmitt, 2002) to video-based tests consisting of multimedia scenarios (Lievens, Buyse, & Sackett, 2005; Olson-Buchanan et al., 1998; Weekley & Jones, 1997). SJTs can, furthermore, vary in terms of their response instructions. McDaniel and Nguyen (2001) identified two categories of response instructions, namely a knowledge response instruction and a behavioral tendency response instruction. Knowledge response instructions typically ask respondents to rate the effectiveness of responses. Behavioral tendency response instructions ask respondents to select the response they would most likely and/or least likely do.

The Concept of Integrity and Integrity Tests

Now, we will turn to one construct, namely integrity, which the SJT in the present study aims to measure. For the following two reasons, more and more attention is given to integrity during personnel selection and for job performance. First, measures of integrity have shown to be predictive of

organizational outcomes, from theft to job performance (Ones, Viswevaran, & Schmidt, 1993). Second, integrity tests have also shown to predict incrementally over and above measures of cognitive ability (Schmidt & Hunter, 1998). The purpose of the present study was to develop a video-based SJT for integrity, i.e., a SJT based on video scenarios intended to measure the construct of integrity. We collected field data in a multi-ethnical setting during Dutch police officer selection. The construct validity of the SJT was, therefore, examined for both the ethnic majority and the ethnic minority group. The largest ethnic minority groups in The Netherlands are from the Dutch Antilles, Morocco, Surinam, and Turkey.

Integrity is difficult to define and appears to consist of various sub-dimensions (Jones, Brasher, & Huff, 2002; Van Iddekinge, Taylor, & Eidson, 2005). Often-found examples of sub-dimensions of integrity are honesty, drug avoidance, work values, and customer service. The present study focuses on integrity as defined within the police context (Naeyé, Huberts, Van Zweden, Busato, & Berger, 2004, p. 19):

> "Police integrity refers to whether the performance in police jobs is in accordance with the applicable values, norms, and the rules that are involved. Values are moral principles or standards, such as legitimacy and brotherhood, which should be of importance during decision-making. Norms are more concrete and direct. Norms are action rules, which give a clear guidance in what is allowed in a specific situation and what is not."

Violations of integrity at the Dutch police involve, among other things, corruption, fraud and theft, accepting dubious gifts and services, misuse of authority, and misuse of information (Naeyé et al., 2004), which can be viewed as sub-dimensions of police integrity. Because of the impact that these integrity violations may have on the police organization, it is important to determine an applicant's integrity by means of a police officer selection measure.

Broadly, there are two types of integrity tests. Tests using items that focus on attitudes toward theft and other dishonest behaviors are referred to as overt integrity tests, whereas tests developed to assess broad personality traits that predict counterproductive behaviors are referred to as personality-based integrity tests or so-called disguised purpose tests (Ones & Viswesvaran, 1998). In general, integrity tests have been found to positively relate to the Big Five personality dimensions of conscientiousness, agreeableness, and emotional stability, with their relative importance in that order. Further, integrity tests have negligible correlations with cognitive ability and score

differences between ethnic groups on integrity tests have been shown to be very small (Ones & Viswesvaran, 1998).

An Integrity-SJT

The decision to construct an Integrity-SJT, instead of developing an overt or personality-based integrity test, pertains to the advantages of SJTs and, more specifically, video-based SJTs. These are, as mentioned earlier, a high criterion-related validity (Lievens & Sackett, 2006), little adverse impact against ethnic minority applicants, and high realism of the test leading to more reliable respondent reactions (Chan & Schmitt, 1997; Richman-Hirsch et al., 2000). The SJT that was developed for the Dutch police consists of videos of critical situations in each of which one of the above-mentioned police-integrity violations are presented.

Little is known in the literature about SJTs that have been developed in a construct-driven way, besides that SJTs – a posteriori – have generally shown to correlate with cognitive ability, conscientiousness, agreeableness, and emotional stability. We know of one empirical study by Becker (2005) that addressed the development and validation of a SJT aimed to measure integrity. Becker argued that his SJT was based on an explicit, clear definition of integrity and was intended to capture specific integrity values rather than general personality traits or other variables that are related to, but not synonymous with, integrity and that a clear definition of integrity was necessary to explain what was measured. Furthermore, Becker found that his SJT was a valid predictor of moderate magnitude of outcomes in real-world settings, such as promotion, career progress, and status as a team leader. The present study builds on the work of Becker (2005) in order to demonstrate that a construct-driven development of SJTs is indeed possible. We developed a SJT intended to measure integrity and investigated its construct validity. To this end, we investigated the relationship between the SJT score and actual integrity-related variables, instead of examining the relationship between the SJT score and general work-related outcomes, as Becker did.

Overview of Hypotheses

Previous research has shown that scores on integrity tests correlate with several other dimensions. For instance, overt integrity tests have been found to show higher correlations with the Big Five personality dimensions of conscientiousness (observed r = .26), agreeableness (observed r = .23), and emotional stability (observed r = .18) than with the Big Five dimensions of extraversion (observed r = .02) and intellect (observed r = .06; Ones, 1993, in Wanek, 1999).

Next to the fact that integrity tests appear to correlate with the Big Five dimensions conscientiousness, agreeableness, and emotional stability, McDaniel and Nguyen (2001) meta-analytically showed that SJTs in general are also correlated with conscientiousness (observed $\bar{r}$ = .26), agreeableness (observed $\bar{r}$ = .25), and emotional stability (observed $\bar{r}$ = .31). Since both integrity tests and SJTs have shown to be related to conscientiousness, agreeableness, and emotional stability, examining correlations between the present SJT and the three Big Five dimensions, solely, will not give much insight into whether the SJT indeed measures integrity. If, for instance, correlations around .25 are found in the present study, would this mean that the SJT measures integrity or would this mean that the test is yet another multidimensional SJT? Therefore, the SJT's convergent validity is examined by means of the relationship between the SJT score and several integrity-related dimensions, namely the dimension Honesty-Humility of the HEXACO-model, cognitive-distortion sub-dimensions of the 'How-I-Think' questionnaire (HIT questionnaire), and behavioral-referent sub-dimensions of the HIT questionnaire. Also, the discriminant validity of the SJT is investigated with non-integrity-related dimensions, such as cognitive ability and several other non-integrity-related dimensions measured throughout the selection process.

In the following, we will state the hypotheses and the arguments for these hypotheses. The first hypothesis states that scores on the Integrity-SJT are more strongly correlated with other integrity-related dimensions than with non-integrity related dimensions (Hypothesis 1). A dimension that has shown a strong resemblance to the concept of integrity is the sixth factor of the recently introduced personality structure HEXACO (Lee & Ashton, 2004). This sixth factor is labeled Honesty-Humility and is typically described as honesty, fairness, sincerity, modesty, and lack of greed. Lee, Ashton, and De Vries (2005) argued that the dimension Honesty-Humility has a clear conceptual link to integrity, since "both consist of admissions of wrongdoing such as theft, fraud, sabotage, and alcohol and drug abuse" (p. 182). Hence, they investigated the relationship between Honesty-Humility on the one hand and workplace delinquency and scores on an overt integrity test on the other hand; they found correlations of -.47 for workplace delinquency and .53 for integrity. Therefore, we expect that the score on the Integrity-SJT will be substantially correlated to the dimension Honesty-Humility.

The HIT questionnaire is a measure of self-serving cognitive distortions (Barriga, Gibbs, Potter, & Liau, 2001). Self-serving cognitive distortions are inaccurate or biased ways of attending to or conferring meaning upon experiences associated with externalizing behavior. An example of a person

showing self-serving cognitive distortions is someone who has been stealing something from a shop but who blames the shop owner for making stealing possible. Barriga and Gibbs (1996) argued that self-serving cognitive distortions should correlate with measures of antisocial behavior, such as theft, fraud, aggressive behavior, and disobedience. They found a correlation of .54 between scores on the HIT questionnaire and aggressive behavior and a correlation of .46 between scores on the HIT questionnaire and delinquent behavior. Therefore, we expect that the score on the Integrity-SJT will also be substantially correlated to scores on the HIT questionnaire.

The second hypothesis states that scores on the Integrity-SJT will show a small correlations with cognitive ability (Hypothesis 2). Regarding the cognitive loading of SJTs, McDaniel et al. (2001) found an observed mean correlation between cognitive ability and SJTs of .36. However, Ones and Viswesvaran (1998) showed that integrity tests have negligible correlations with cognitive ability. Since it is integrity that the SJT intends to measure, we expect a small correlation between the SJT score and scores on the cognitive ability test.

The third hypothesis states that score differences between the ethnic majority and minority group on the Integrity-SJT are smaller than score differences on the cognitive ability test (Hypothesis 3). One of the reasons why SJTs have gained increasing popularity in recent years is the finding of smaller score differences between the ethnic majority and minority group on the SJT than on cognitive ability tests (e.g., Clevenger et al., 2001; Motowidlo et al., 1990). In a meta-analysis, Nguyen, McDaniel, and Whetzel (2005) showed that score differences on SJTs between Blacks and Whites were around .38 *SD* favoring Whites. O'Connell et al. (2007) also found a score difference between Blacks and Whites of .38 *SD*. Furthermore, Ones and Viswesvaran (1998) showed that score differences between ethnic groups on integrity tests are very small (all below .15 *SD*). Therefore, we expect the score difference between the ethnic majority and minority group to be smaller than on the cognitive ability test.

6.2 Method

Sample and Procedure

Data came from ethnic majority and ethnic minority applicants who applied for a position at the Police Academy of The Netherlands in the period from March 2005 until August 2006. The dataset consisted of 1,696 applicants

(68% male; M_{age} = 24.39, *SD* = 6.77), of which 1,371 were ethnic majority applicants (67% male, M_{age} = 24.39, *SD* = 6.84) and 189 were ethnic minority applicants (70% male, M_{age} = 24.41, *SD* = 6.33). Data of 136 applicants were incomplete and, therefore, were removed from the dataset. The ethnic minority applicants were from the Dutch Antilles, Morocco, Surinam, and Turkey. Applicants who are interested in a job as police officer first apply to the local police force where they want to work after training. For the selection procedure, the local police forces routinely send all applicants to the national Police Center for Competence Assessment and Monitoring (CCM). During a requirement check at the CCM, the following minimal criteria are checked on the basis of an application form: minimal age (16 years), Dutch nationality, possession of a swimming diploma, no criminal record, and possession of a school diploma (minimal level is preparatory vocational education level B [VBO-B]). Applicants in the selection process go through two stages. The present study focuses on the second stage, during which applicants go through a cognitive ability test, a personality questionnaire, an assessment center (AC) assignment, and an employment interview. The psychologist who conducts the interview also is the one who writes the final employment recommendation to the local police force. For the employment recommendation, the test results of the personality questionnaire, the AC ratings, and the employment interview ratings are used. Next to the final recommendation, the final dossier to the local police force includes test scores of the cognitive ability test.

Measures

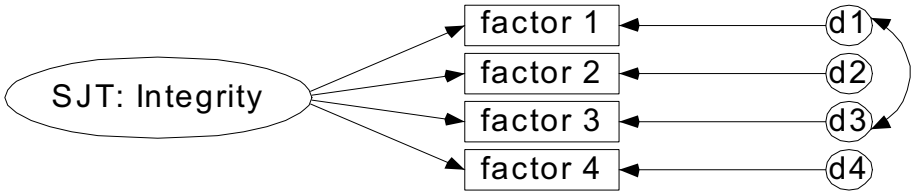
SJT for Integrity

The Integrity-SJT included scenarios representing interpersonal situations, which is a feature of SJTs in general. We used an approach analogous to other studies (see, e.g., Weekley & Jones, 1997) for its development (for an example of an SJT-item, see Appendix C). First, we collected realistic critical incidents regarding interactions between police officers and civilians or among police colleagues from fifteen experienced police officers (both policemen and policewomen; both ethnic majority and minority police officers; police experts had around 15 years of police work experience). All incidents focused on integrity violations and potential reactions to these violations. For example, several incidents dealt with resisting fraudulent people or situations. Second, critical incidents that were similar were grouped and scenarios were written about each of these groups of critical incidents. The fifteen experienced police officers who had been interviewed to collect the critical incidents also checked the scenarios for realism. At the same time, with the help of these experienced police officers, four response options were derived for each scenario. This resulted in fourteen SJT items (a scenario including its

four response options will be named 'item') that were pilot tested in a written version of the test. Third, after examining the descriptives and the factor-analytic results of the pilot-study data (N = 228, 72% male, M_{age} = 24.08; SD = 6.78), three of the fourteen SJT items were eliminated. Fourth, both professional actors and police officers were trained to act in scenarios. After that, the scenarios were videotaped in a professional manner. Several police officers were present during the video-shoot and were asked, again, to assess the scenarios in terms of their realism. Finally, a panel of experts was asked to fill out the video-based SJT in order to develop a scoring key. The expert panel consisted of 50 experienced police officers with on average 14.06 years of work experience (SD = 6.38) and with different ethnic backgrounds, namely 10 ethnic majority experts, 10 Antillean experts, 10 Moroccan experts, 10 Surinamese experts, and 10 Turkish experts. Each response option had to be evaluated on its effectiveness given the situation presented in the scenario. Agreement among the experts in effectiveness ratings was generally satisfactory (mean intraclass correlation [ICC] = .70), both within ethnic groups (mean ICC = .69) and between ethnic groups (mean ICC = .69). Since agreement among experts was satisfactory, the scoring key was set at the modus of the total expert group. The absolute difference between the scoring key of a given item response option and the applicant response formed the applicant score, varying from 4 (largest difference between expert and applicant response) to 0 (no difference between expert and applicant response). The applicant score was subtracted from 4, in order to have an intuitively logical range from 0 (lowest possible score) to 4 (highest possible score). In its final form, the video-based SJT consists of short, videotaped scenarios of key integrity issues that police officers are likely to encounter with civilians or with police colleagues. A narrator introduces each scenario. Per SJT item, the scene freezes at an important point and the applicant has to answer the responses related to the scene presented. The eleven items have four response options each. Applicants have to evaluate each response option in terms of its effectiveness within the given situation. This response instruction generally is known as a knowledge response instruction.

The SJT structure was analyzed with Structural Equation Modeling (SEM) using Amos 6.0 (Arbuckle, 2005). Figure 1 shows the best fitting model (χ^2 [df = 1] = .01, ns; TLI = 1.00; CFI = 1.00; RMSEA = .00). The four sub-factors turned out to represent meaningful clusters of response options (i.e., Factor 1 represented applicant scores on the response option that can generally be described as "It is alright for this time.", Factor 2: "It is not permitted!!" [in a stern way], Factor 3: "These are the rules, so it is not allowed." [in a more friendly way], and Factor 4: "It is not allowed and I have to report it to the supervisor!"), which all loaded significantly ($.33 < \beta < .58$, $p < .001$) on one

general SJT factor. Intercorrelations between the four clusters of response options varied between -.03, *ns* and .21, $p < .01$ ($\bar{r} = .13$). The error terms of Factor 1 and Factor 3 appeared to be somewhat negatively correlated ($r = -.18$, $p < .001$). As the clusters of response options were not conceptually meaningful in terms of Integrity sub-dimensions, further analyses were conducted with the general SJT score. The internal consistency of the SJT was .69.



Note. Each factor reflects a cluster of response options.

Figure 1. SJT model with 4 underlying sub-factors

Selection Measures

The selection procedure consisted of a cognitive ability test, a personality questionnaire, an AC, and an employment interview. The scores on the personality questionnaire, the AC, and the employment interview were integrated into a final employment recommendation.

Cognitive Ability Test. The Police Intelligence Test (PIT; Rijks Psychologische Dienst, 1975) is a cognitive ability test and consists of 107 items divided over six subtests: Verbal Comprehension, Inductive Reasoning, Numerical Reasoning, Word Fluency, Spatial ability, and Picture Arrangement. The time limit is 51 minutes. Applicants completed the PIT in Dutch. Prior research by Lem and Van Doorn (2000) indicated alpha reliabilities varying from .69 for Series of Numbers, to .87 for Folding Figures. The correlations between the subscales varied from .32 to .57. A study by Van der Maesen (1992) showed corrected predictive validity coefficients of .39 and .46 ($N = 162$).

Personality Questionnaire. To measure the Big Five factors Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect, the Police Personality Questionnaire (PPV; Van Leeuwen, 2000) was used. The applicants completed the PPV in Dutch. A recent progress report by Klinkenberg and Van Leeuwen (2003) indicated alpha reliabilities varying

from .72 for Conscientiousness, to .78 for Intellect. Correlations between the scales all are lower than .60. Comparison with NEO-PI-R showed observed construct validity coefficients between .17 and .58 (N = 160). A study by Lem and Van Doorn (2000) showed observed predictive validity coefficients between .15 and .43 (N = 61).

Assessment Center (AC). A role-play exercise was utilized, in which an assessor and an actor independently made ratings on a 7-point Likert-scale ranging from 1 (extremely weak) to 7 (excellent), on each of the following seven dimensions: Communication Skills, Social Skills, Empathy, Initiative, Stress Tolerance, Authority, and Decisiveness. Prior research (De Meijer, Born, Terlouw, & Van der Molen, 2006) showed that interrater reliabilities ranged from .82 to .88 (N = 198) and principal component analysis with varimax rotation yielded two factors, Agency and Communion (in accordance with Wiggins and Trapnell, 1996), which together explained 77% of the variance. As a measure of Agency, the average rating across the dimensions of Authority, Decisiveness, Initiative, Communication Skills, and Stress Tolerance was used ($\bar{r}$ = .59; α = .87). As a measure of Communion, the average rating of the dimensions Social Skills and Empathy was used ($\bar{r}$ = .77; α = .87). The reliability of the difference (r_{diff}) between scores on Agency and Communion was .78.

Employment Interview. The interview questions were focused on evaluating behavior on the following eight dimensions: Communication Skills, Social Skills, Flexibility, Stress Tolerance, Emotional Stability, Tolerance Towards Others, Integrity, and Self-Understanding. A single interviewer conducted the interview. The interviews were semi-structured and behaviorally based, with one behaviorally anchored 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent) for each of the eight dimensions. In the present study, we focused on the dimension Integrity (for a definition, see Appendix D).

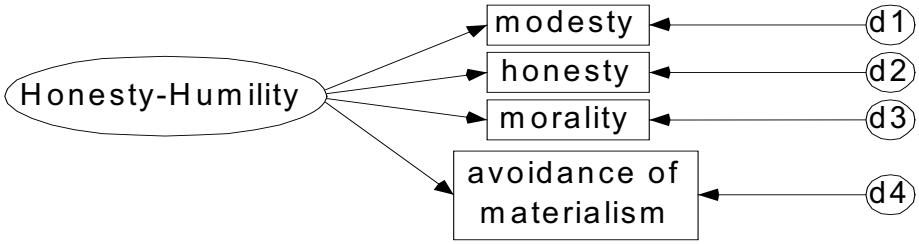
Final Employment Recommendation. The final recommendation as to whether an applicant is fit for a job as police officer was based on results from the personality test (PPV), the AC, and the employment interview. These scores were integrated into an employment recommendation. The dimensions in the final recommendation were: Communication Skills, Social Skills, Empathy, Initiative, Flexibility, Stress Tolerance, Authority, Decisiveness, Tolerance Towards Others, Integrity, and Self-Understanding (for definitions, see Appendix B). A 7-point Likert scale ranging from 1 (extremely weak) to 7 (excellent) was used to evaluate the behavior on the eleven dimensions. Prior research (De Meijer et al., 2006) conducting

principal component analysis with varimax rotation yielded three employment-recommendation factors, Agency, Communion, and Socio-Cultural Awareness, which altogether explained 67% of the variance. As a measure of Agency, the average rating across the dimensions Authority, Decisiveness, Initiative, Communication Skills, Stress Tolerance, and Flexibility was used ($\bar{r}$ = .48; α = .85). As a measure of Communion, the dimensions Social Skills and Empathy, were used ($\bar{r}$ = .66; α = .79) and for Socio-Cultural Awareness the dimensions ($\bar{r}$ = .39; α = .65), Tolerance Towards Others, Integrity, and Self-Understanding. The reliability of the difference (r_{diff}) between scores on Agency and Communion is .51, r_{diff} between scores on Agency and Socio-Cultural Awareness is .58, and r_{diff} between scores on Communion and Socio-Cultural Awareness is .57. In the present study, we focused on the factors Agency and Communion and, separately, on the dimension Integrity.

Other Integrity Measures

At the end of the selection procedure, a sub-sample of 204 applicants (of which 50 ethnic minority applicants) was available to participate in an in-depth Honesty-Humility interview and to fill out the HIT questionnaire.

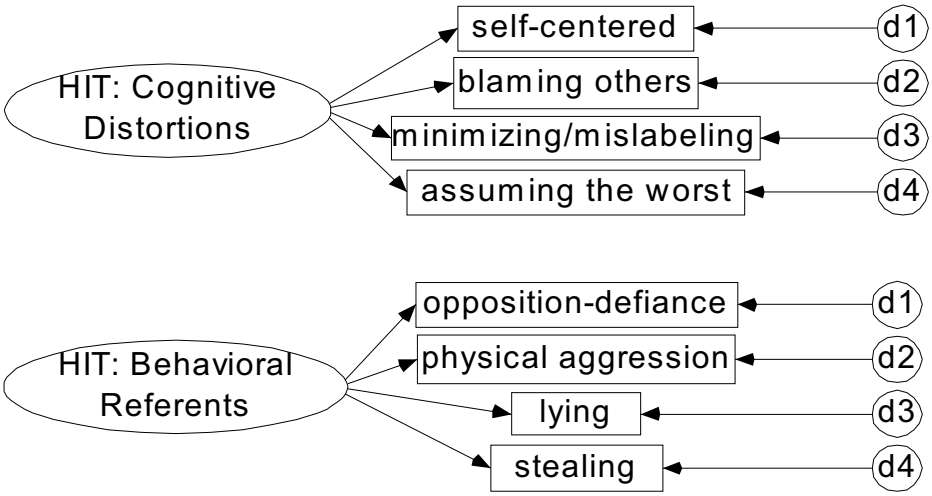
In-Depth Interview. The in-depth interview was built around the sixth factor of the HEXACO-model (Lee & Ashton, 2004) 'Honesty-Humility' and its four sub-dimensions: Modesty, Honesty, Morality, and Avoidance of Materialism (for definitions, see Appendix D). The interviews were semi-structured and behaviorally based. The interviewer and an assessor, who was present during the interview, independently made ratings on a 7-point Likert-scale ranging from 1 (extremely weak) to 7 (excellent), on the factor Honesty-Humility and each of the four sub-dimensions. Interrater reliabilities ranged from .63 to .78 (N = 203). The interview ratings were used for further analyses. Confirmatory factor analyses were conducted on several models. The model presented in Figure 2 showed the best fit, which can be indicated as a good fit of the data to the model (χ^2 [df = 2] = 1.06, *ns*; TLI = 1.00; CFI = 1.00; RMSEA = .01). All sub-dimensions loaded significantly (.45 < β < .83, p < .001) on the Honesty-Humility factor. Intercorrelations between the four sub-dimensions varied between .17, p < .05 and .45, p < .01 ($\bar{r}$ = .34). Both the Honesty-Humility factor and its sub-dimensions were used in further analyses.



Note. Each factor reflects a sub-facet of Honesty-Humility.

Figure 2. In-depth interview model with 4 underlying sub-facets

'How-I-Think' Questionnaire (HIT questionnaire). To measure applicants' cognitive distortions, the Dutch translation (translated from English by Utrecht University, The Netherlands) of the HIT questionnaire (Barriga et al., 2001) was used. The HIT questionnaire was developed to measure two broad dimensions, namely Cognitive Distortions and Behavioral Referents, each consisting of four sub-dimensions. Cognitive Distortions consists of the sub-dimensions Self-Centered, Blaming Others, Minimizing/Mislabeling, and Assuming the Worst. Behavioral Referents consists of the sub-dimensions Opposition-Defiance, Physical Aggression, Lying, and Stealing (for definitions, see Appendix D). The alpha reliability of the dimension Cognitive Distortions was .90 and of the dimension Behavioral Referents was .89, based on the present sample. The alpha reliabilities of the sub-dimensions varied from .70 for Blaming Others to .79 for Stealing. Confirmatory factor analyses were conducted on several models. The models presented in Figure 3 showed the best fit, which can be indicated as a good fit of the data to the models (χ^2 [df = 2] = 4.41, $p < .05$, and 0.83, *ns*; TLI = .94 and 1.00; CFI = .99 and 1.00; RMSEA = .05 and .00). All Cognitive Distortions sub-dimensions loaded significantly ($.80 < \beta < .90$, $p < .001$) on the Cognitive Distortions dimension. All Behavioral Referents sub-dimension loaded significantly ($.75 < \beta < .90$, $p < .001$) on the Behavioral Referents dimension. Intercorrelations between the sub-dimensions of the Cognitive Distortions dimension varied between .73, $p < .01$ and .75, $p < .01$ ($\bar{r} = .74$). Intercorrelations between the sub-dimensions of the Behavioral Referents dimension varied between .60, $p < .01$ and .62, $p < .01$ ($\bar{r} = .61$). The two dimensions as well as their sub-dimensions were used in further analyses.



Note. Each factor reflects a sub-dimension of either Cognitive Distortions or Behavioral Referents.

Figure 3. Two dimensions of the How-I-Think-questionnaire, each containing 4 underlying sub-dimensions

Analyses

Preliminary Analyses

Because response styles can affect answers on questionnaires (e.g., Van Herk, Poortinga, & Verhallen, 2004), structural equivalence (i.e., absence of bias) of all measures was checked across ethnic groups before conducting further analyses. In accordance with Van Herk et al. (2004), structural equivalence across cultures is interpreted as follows: A test measures the same trait cross-culturally, but not necessarily on the same quantitative scale. Using Amos 6.0 (Arbuckle, 2005), no differences between factor structures of all measures were found between the ethnic majority group and the minority group (for detailed information, please contact the first author).

Main Analyses

Correlations among the various integrity-related measures and (sub-) dimensions were calculated to examine the convergent validity of the SJT. Correlations between the SJT and various non-integrity related dimensions were calculated to investigate its discriminant validity. Correlations were calculated for the ethnic majority group and for the ethnic minority group, separately. Fisher's logarithmic transformation (for the formula, see

Lindeman, Merenda, & Gold, 1980) of the correlation coefficients was used to test for significant differences between correlations of the ethnic majority and minority group.

Standardized effect sizes (*d* values) between the means of the ethnic majority and the ethnic minority group were computed to get an indication of the magnitude of the group differences on the various instruments irrespective of sample size. Positive *d* values indicate higher mean scores for the majority group and negative *d* values indicate higher mean scores of a minority group. Although effect sizes can theoretically range between positive and negative infinity, Cohen (1988) suggests that effect sizes of about .20 in magnitude are small, around .50 are medium, and above .80 are large. To compute *d* values, observed differences on dimension scores were used that were uncorrected for age, gender, and education. Corrected *d* values only differed marginally (about .01 *SD*) from uncorrected *d* values.

6.3 Results

First, we expected that scores on the Integrity-SJT would be more strongly correlated with other integrity-related tests than with non-integrity related tests (Hypothesis 1). We examined the correlations for the ethnic majority group and for the ethnic minority group, separately. The correlations between all measures are reported in Tables 1 (convergent validity) and 2 (discriminant validity). All reported correlations are observed correlations.

On the convergent personality dimensions Agreeableness, Conscientiousness, and Emotional Stability, the correlations with the SJT for the majority group were .16 ($p < .01$), .17 ($p < .01$), and .22 ($p < .01$), respectively. For the ethnic minority group the correlations were, resp., .10 (*ns*), .08 (*ns*), and .17 ($p < .05$). Furthermore, Integrity was measured during the employment interview and the final employment recommendations comprised a score on Integrity. For the ethnic majority group, correlations between these two integrity dimensions and the SJT were both .16 ($p < .01$). For the ethnic minority group, correlations were .04 (*ns*) for the employment interview and .06 (*ns*) for the final employment recommendation. No significant differences in correlations between the ethnic majority and minority group were found.

Table 1

Convergent Validity: Correlations Among the various Integrity-Related Measures and (Sub-) Dimensions: The SJT, the Personality Dimensions Agreeableness, Conscientiousness, and Emotional Stability, the Employment-Interview Dimension Integrity, the Final-Recommendation Dimension Integrity, the In-Depth Interview, and the How-I-Think Questionnaire

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Integrity-SJT	-	0.10	0.08	0.17*	0.04	0.06	0.21	0.04	0.13	0.30
2. Personality: Agreeableness	0.16**	-	0.39**	0.28**	-0.05	-0.02	-0.11	0.08	0.10	0.05
3. Personality: Conscient.ness	0.17**	0.37**	-	0.31**	-0.07	-0.05	-0.21	0.07	0.09	-0.01
4. Personality: Emot. Stability	0.22**	0.31**	0.28**	-	0.07	0.05	0.02	0.02	0.35*	-0.07
5. Employment Interview	0.16**	0.06	0.08**	0.04	-	*0.98***	0.16	0.21	0.20	0.03
6. Final Recommendation	0.16**	0.05	0.09**	0.05	*0.95***	-	0.19	0.26	0.18	0.09
7. In-Depth Interview	0.27**	0.05	-0.04	-0.05	0.18*	0.15	-	0.60**	*0.25*	0.50**
8. In-Depth Interview: Modesty	0.24**	0.06	0.03	-0.09	0.06	0.04	0.60**	-	-0.06	0.10
9. In-Depth Interview: Honesty	0.11	-0.04	-0.08	0.15	0.16	0.13	*0.68***	0.25**	-	*0.05*
10. In-Depth Interview: Morality	0.23**	0.21**	0.07	0.10	0.22**	0.22**	0.64**	0.32**	*0.40***	-
11. In-Depth Interv.: Av. of Mat.	0.10	0.01	0.01	-0.11	0.14	0.14	0.74**	0.39**	*0.51***	0.50**
12. HIT: Cognitive Distortions	-0.34**	-0.34**	-0.23**	-0.36**	-0.23**	-0.20*	-0.10	-0.05	-0.05	-0.26**
13. HIT: Self-Centered	-0.34**	-0.25**	-0.19*	-0.34**	-0.16	-0.13	-0.11	-0.05	-0.13	-0.23**
14. HIT: Blaming Others	-0.31**	-0.31**	-0.22**	-0.35**	-0.24**	-0.23**	-0.03	-0.06	0.03	-0.17*
15. HIT: Minimizing/Mislabeling	-0.30**	-0.27**	-0.23**	-0.28**	-0.22**	-0.20*	-0.12	-0.06	-0.06	-0.28**
16. HIT: Assuming the Worst	-0.26**	-0.38**	-0.16*	-0.31**	-0.21*	-0.17*	-0.08	0.00	-0.01	-0.23**
17. HIT: Behavioral Referents	-0.34**	-0.34**	-0.23**	-0.36**	-0.23**	-0.20*	-0.10	-0.05	-0.05	-0.26**
18. HIT: Opposition-Defiance	-0.32**	-0.29**	-0.20*	-0.33**	-0.24**	-0.22**	-0.03	0.00	-0.04	-0.20*
19. HIT: Physical Aggression	-0.24**	-0.30**	-0.17*	-0.35**	-0.21**	-0.22**	-0.10	*-0.05*	-0.02	-0.23**
20. HIT: Lying	-0.30**	-0.27**	-0.22**	*-0.32***	-0.14	-0.10	-0.17*	-0.06	-0.13	-0.28**
21. HIT: Stealing	-0.32**	-0.33**	-0.20*	-0.25**	-0.20*	-0.17*	-0.02	-0.06	*0.02*	-0.17*

Table 1 – continued

Measure	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.
1. SJT	0.15	-0.43**	0.42**	-0.37*	-0.37*	-0.30	-0.43**	-0.47**	-0.27	-0.40**	-0.33*
2. Personality: Agreeableness	-0.19	-0.43**	-0.25	-0.41**	-0.33*	-0.50**	-0.44**	-0.34*	-0.42**	-0.32*	0.41**
3. Personality: Conscient.ness	-0.23	-0.15	-0.04	0.15	-0.13	-0.19	-0.15	-0.20	-0.11	-0.07	-0.14
4. Personality: Emot. Stability	0.05	-0.15	0.02	-0.25	-0.07	-0.22	-0.15	-0.18	-0.19	*0.03*	-0.21
5. Employment Interview	0.28	-0.15	-0.25	0.00	-0.12	-0.13	-0.14	-0.17	-0.15	-0.15	0.00
6. Final Recommendation	0.24	-0.18	-0.29	-0.01	-0.16	-0.15	-0.18	-0.16	-0.16	-0.22	-0.03
7. In-Depth Interview	0.79**	-0.25	-0.23	0.17	-0.22	-0.25	-0.25	-0.10	-0.31*	-0.14	-0.29
8. In-Depth Interview: Modesty	0.42**	-0.37*	-0.32*	-0.29	-0.39**	-0.29	-0.37*	-0.19	*-0.43***	-0.33*	-0.28
9. In-Depth Interview: Honesty	*0.20*	-0.09	0.10	-0.15	-0.09	-0.17	-0.09	-0.10	-0.06	0.12	*-0.32**
10. In-Depth Interview: Morality	0.34*	-0.24	-0.18	-0.08	-0.25	-0.31*	-0.23	-0.22	-0.21	-0.12	-0.26
11. In-Depth Interv.: Av. of Mat.	-	-0.13	-0.19	-0.06	-0.09	-0.09	-0.12	-0.10	-0.11	-0.06	-0.15
12. HIT: Cognitive Distortions	-0.10	-	0.79**	0.86**	0.92**	0.86**	1.00**	0.85**	0.88**	0.82**	0.83**
13. HIT: Self-Centered	-0.14	0.89**	-	0.50**	0.68**	0.51**	*0.78***	*0.60***	0.60**	0.81**	0.58**
14. HIT: Blaming Others	-0.01	0.88**	0.67**	-	0.73**	0.71**	0.86**	0.78**	0.79**	0.58**	0.80**
15. HIT: Minimizing/Mislabeling	-0.11	0.91**	0.77**	0.73**	-	0.75**	0.92**	0.73**	0.82**	0.77**	0.77**
16. HIT: Assuming the Worst	-0.10	0.88**	0.70**	0.75**	0.74**	-	0.87**	0.82**	0.82**	0.63**	0.69**
17. HIT: Behavioral Referents	-0.10	1.00**	*0.88***	0.88**	0.91**	0.89**	-	0.85**	0.88**	0.83**	0.83**
18. HIT: Opposition-Defiance	-0.04	0.90**	*0.79***	0.78**	0.81**	0.86**	0.90**	-	0.72**	0.59**	0.61**
19. HIT: Physical Aggression	-0.12	0.87**	0.71**	0.82**	0.80**	0.77**	0.87**	0.73**	-	0.57**	0.67**
20. HIT: Lying	-0.16	0.85**	0.88**	0.67**	0.78**	0.70**	0.86**	0.68**	0.64**	-	0.58**
21. HIT: Stealing	-0.01	0.85**	0.67**	0.82**	0.76**	0.79**	0.85**	0.76**	0.61**	0.63**	-

Note. HIT means the How-I-Think-questionnaire. The total sample consists of 1,696 applicants. Values of ethnic majority applicants are below the diagonal and values of ethnic minority applicants (N = 189 in total sample and N = 50 for the in-depth Integrity-interview and the HIT questionnaire) are above the diagonal. Significantly different correlations (tested with Fisher's logarithmic transformation) are in *italics*.

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

Table 2

Disciminant Validity: Correlations Between the SJT and various Non-Integrity Related Dimensions: The Personality Dimensions Extraversion and Intellect, the AC Factors Agency and Communion, and the Final-Recommendation Factors Agency and Communion

Measure	1.	2.	3.	4.	5.	6.	7.
1. Integrity-SJT	-	-0.03	0.14	0.15	-0.02	0.27**	0.10
2. Personality: Extraversion	0.09**	-	0.57**	*0.27***	0.18*	*0.32***	0.18*
3. Personality: Intellect	0.21**	0.50**	-	0.26**	0.23**	0.29**	0.15
4. AC: Agency	0.15**	*0.10***	0.14**	-	0.51**	0.82**	0.41**
5. AC: Communion	0.06*	0.01	0.02	0.52**	-	0.34**	0.74**
6. Final Recommendation: Agency	0.20**	*0.15***	0.23**	0.84**	0.41**	-	0.47**
7. Final Recommendation: Communion	0.07*	0.09**	0.06	0.45**	0.77**	0.49**	-

Note. The total sample consists of 1,696 applicants. Values of ethnic majority applicants are below the diagonal and values of ethnic minority applicants (N = 189) are above the diagonal. Significantly different correlations (tested with Fisher's logarithmic transformation) are in *italics*.
* $p < .05$, ** $p < .01$.

For the ethnic majority group, correlations of the SJT with the sixth factor of the HEXACO-model, Honesty-Humility, and its sub-dimensions were .27 ($p < .01$) for general Honesty-Humility and, on its sub-dimensions, varying between .11 (*ns*) for Honesty and .24 ($p < .01$) for Modesty. For the ethnic minority group, the correlations that were found were .21 (*ns*) for general Honesty-Humility and, concerning its sub-dimensions, varying between .04 (*ns*) for Modesty and .30 (*ns*) for Morality. For the HIT-dimensions Cognitive Distortions and Behavioral Referents and its sub-dimensions, the correlations for the ethnic majority group were both -.34 ($p < .01$), varying between -.24 ($p < .01$) on Physical Aggression and -.34 ($p < .01$) on Self-Centered. For the ethnic minority group, the correlations were both -.43 ($p < .01$), varying from -.27 on Physical Aggression to -.47 ($p < .01$) on Opposition-Defiance. Again, no significant differences in correlations between the ethnic majority and minority group were found.

In sum, regarding the convergent-validity evidence, the correlations were rather low between the SJT and the personality dimensions Agreeableness, Conscientiousness, and Emotional Stability for both the ethnic majority and minority group. On the employment interview and the final employment recommendation, the correlations between the SJT and Integrity were also rather low, especially for the ethnic minority group. On the HEXACO-factor Honesty-Humility and on the HIT questionnaire, however, correlations were moderate in size for both the ethnic majority and minority group.

Regarding the discriminant-validity results, the correlations between the SJT and the non-integrity related dimensions were all below .21 for the ethnic majority group and below .27 for the ethnic minority group. Especially on the personality dimension Extraversion and on the AC- and final-recommendation factor Communion, correlations were low for both the ethnic majority group ($r = .09$, $p < .01$; $r = .06$, $p < .05$; and $r = .07$, $p < .05$, resp.) and the ethnic minority group ($r = -.03$, *ns*; $r = -.02$, *ns*; and $r = .10$, *ns*, resp.). On the personality dimension Intellect ($r = .21$, $p < .01$, for the ethnic majority group and $r = .14$, *ns,* for the ethnic minority group), on the AC-factor Agency ($r = .15$, $p < .01$, for the ethnic majority group and $r = .15$, *ns,* for the ethnic minority group), and the final-recommendation factor Agency ($r = .20$, $p < .01$, for the ethnic majority group and $r = .27$, $p < .01$, for the ethnic minority group) the correlations were somewhat higher for both groups than on the other non-integrity related dimensions.

In conclusion, the correlations on the integrity-related (sub-) dimensions, especially on Honesty-Humility and the HIT questionnaire, were higher than on the non-integrity related dimensions. These findings support Hypothesis 1

and demonstrate that the SJT indeed measures Integrity for both ethnic groups, as no significant differences in correlations between the ethnic majority and minority group were found.

Second, we expected that scores on the Integrity-SJT would show only small correlations with cognitive ability (Hypothesis 2). Table 3 reports the correlations between the SJT and the cognitive ability sub-tests separately for the ethnic majority and minority group. All reported correlations are observed correlations. For the ethnic majority group, the correlations varied between -.01 (*ns*) on Spatial Ability to .08 (*ns*) on Picture Arrangement. For the ethnic minority group, the correlations varied between .08 (*ns*) on Numerical Reasoning to .19 ($p < .05$) on Inductive Reasoning. The mean correlation between the SJT and cognitive ability was .04 for the ethnic majority group and .13 for the ethnic minority group. Although the correlation between the SJT and cognitive ability was somewhat higher for the ethnic minority group than for the majority group, the general conclusion is that both correlations are quite low, which supports Hypothesis 2. Again, no significant differences in correlations between the ethnic majority and minority group were found.

Finally, we expected that score differences between the ethnic majority and minority group on the Integrity-SJT would be smaller than score differences on the cognitive ability test (Hypothesis 3). Table 4 shows *d* values of all measures and dimensions used. We found a *d* value of .38 *SD* for the score difference between the ethnic majority and the ethnic minority group on the SJT, favoring the ethnic majority group. The mean score difference on the cognitive ability test was .48 *SD*, varying from .38 *SD* for Spatial Ability to .64 *SD* on Word Fluency. The score difference on the SJT, thus, appeared to be smaller than the score difference on the cognitive ability test. Therefore, Hypothesis 3 was supported.

Table 3

Correlations between the SJT and the Cognitive Ability Test

Measure	1.	2.	3.	4.	5.	6.	7.
1. Integrity-SJT	-	0.18*	0.19*	0.08	0.13	0.09	0.10
2. Cognitive Ability: Verbal Comprehension	0.05	-	0.56**	0.50**	0.48**	0.42**	0.41**
3. Cognitive Ability: Inductive Reasoning	0.03	0.41**	-	0.52**	0.35**	0.57**	0.34**
4. Cognitive Ability: Numerical Reasoning	0.06	0.35**	0.44**	-	0.41**	0.48**	0.31**
5. Cognitive Ability: Word Fluency	0.04	0.37**	0.32**	0.39**	-	0.34**	0.17
6. Cognitive Ability: Spatial Ability	-0.01	0.30**	0.51**	0.36**	0.31**	-	0.37**
7. Cognitive Ability: Picture Arrangement	0.08	0.24**	0.33**	0.18**	0.21**	0.21**	-

Note. The total sample consists of 1,696 applicants. Values of ethnic majority applicants are below the diagonal and values of ethnic minority applicants (N = 189) are above the diagonal. Significantly different correlations (tested with Fisher's logarithmic transformation) are in *italics*.

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

Table 4

Descriptives and Comparison Between Ethnic Majority and Minority Group

Measure	Mean and *SD*				Comparison		
	Majority group		Minority group		Levene's test	*t* test	*d*
1. Integrity-SJT	3.11	0.24	3.02	0.24	0.22	4.81**	0.38
2. Cognitive Ability: Verbal Comprehension	9.95	2.44	8.79	2.78	4.57*	4.29**	0.46
3. Cognitive Ability: Inductive Reasoning	10.31	2.58	8.93	3.01	6.37*	4.76**	0.51
4. Cognitive Ability: Numerical Reasoning	8.11	3.83	6.24	3.68	0.22	4.96**	0.48
5. Cognitive Ability: Word Fluency	10.64	3.96	8.07	3.43	5.42*	7.39**	0.64
6. Cognitive Ability: Spatial Ability	21.42	4.19	19.79	4.71	3.79	3.87**	0.38
7. Cognitive Ability: Picture Arrangement	7.82	2.00	7.02	2.06	1.53	4.03**	0.39
8. Personality: Extraversion	5.72	2.69	5.10	2.66	0.18	2.95*	0.23
9. Personality: Agreeableness	5.43	2.67	5.50	2.88	2.07	-0.36	-0.03
10. Personality: Conscientiousness	5.64	2.74	6.21	2.66	2.83	-2.69*	-0.21
11. Personality: Emotional Stability	5.53	2.78	5.97	2.79	0.00	-2.04*	-0.16
12. Personality: Intellect	5.75	2.71	5.21	3.03	8.35*	2.29*	0.20
13. AC: Agency	4.15	0.81	3.65	0.83	0.04	6.99**	0.60
14. AC: Communion	4.04	1.04	3.68	1.14	3.56	3.87**	0.34
15. Employment Interview: Integrity	4.44	0.58	4.28	0.55	13.86**	3.15*	0.28
16. Final Recommendation: Agency	4.19	0.69	3.71	0.68	0.26	7.70**	0.69
17. Final Recommendation: Communion	4.30	0.81	3.98	0.86	0.31	4.34**	0.39
18. Final Recommendation: Integrity	4.43	0.59	4.28	0.54	15.84**	2.97*	0.26
19. In-Depth Interview	4.65	0.85	4.41	0.90	0.21	1.63	0.28
20. In-Depth Interview: Modesty	4.43	1.01	4.02	1.07	0.94	2.33*	0.40

21. In-Depth Interview: Honesty	4.72	0.98	4.75	0.99	0.24	-0.17	-0.03
22. In-Depth Interview: Morality	4.91	0.86	4.70	0.90	0.28	1.36	0.24
23. In-Depth Interview: Avoidance of Materialism	4.45	1.13	4.50	1.02	0.42	-0.26	-0.05
24. HIT: Cognitive Distortions	1.70	0.39	1.82	0.39	0.20	-1.84	-0.31
25. HIT: Self-Centered	1.73	0.48	1.81	0.48	0.01	-1.03	-0.17
26. HIT: Blaming Others	1.74	0.44	1.89	0.47	0.47	-1.89	-0.33
27. HIT: Minimizing/Mislabeling	1.65	0.44	1.75	0.43	0.04	-1.39	-0.23
28. HIT: Assuming the Worst	1.67	0.36	1.82	0.42	3.29	-2.33*	-0.39
29. HIT: Behavioral Referents	1.70	0.38	1.82	0.39	0.28	-1.87	-0.31
30. HIT: Opposition-Defiance	1.63	0.39	1.83	0.41	0.32	-3.04*	-0.50
31. HIT: Physical Aggression	1.87	0.48	2.08	0.50	0.16	-2.45*	-0.43
32. HIT: Lying	1.83	0.50	1.93	0.53	0.73	-1.18	-0.20
33. HIT: Stealing	1.47	0.39	1.45	0.39	0.07	0.28	0.05

Note. HIT means the How-I-Think-questionnaire. The total sample consists of 1,696 applicants (ethnic minority applicants: $N = 189$ in total sample and $N = 50$ for the in-depth Integrity-interview and the HIT questionnaire). Scores on the SJT varied between 0 and 4. Scores on the cognitive ability test varied between 1 and 15 on Verbal Comprehension and Inductive Reasoning, between 1 and 13 on Numerical Reasoning, between 1 and 25 on Word Fluency, between 1 and 29 on Spatial Ability, and between 1 and 10 on Picture Arrangement. Scores on the personality questionnaire varied between 1 and 10. Scores on the AC, the employment interview, the final recommendation, and the in-depth Integrity interview varied between 1 and 7. Scores on the HIT varied between 1 and 6. A positive *d*-value means the ethnic majority group is scoring higher.

* $p < .05$, ** $p < .001$.

6.4 Discussion

In a field study conducted in a multi-ethnical setting at the Dutch police, we examined the construct validity of a video-based situational judgment test (SJT) measuring integrity. We investigated convergent and discriminant validity of the SJT, including correlations between the SJT score and the score on a cognitive ability test, and score differences on all measures between the ethnic majority group and the ethnic minority group. Convergent- and discriminant-validity coefficients were calculated for the ethnic majority and minority group, separately. The largest ethnic minority groups in The Netherlands are from the Dutch Antilles, from Morocco, Surinam, and Turkey.

Concerning the first hypothesis, we found support for the construct validity of the SJT. Although the relationships between the SJT and the personality dimensions Agreeableness, Conscientiousness, and Emotional Stability were rather low, as were the relationships between the SJT and the employment-interview Integrity dimension and the final-recommendation Integrity dimension, the correlations between the SJT and Honesty-Humility and between the SJT and the HIT questionnaire were substantial. Furthermore, correlations between the SJT and the integrity-related (sub-) dimensions were higher than between the SJT and the non-integrity-related dimensions. Especially the correlations with the Honesty-Humility sub-dimension Morality (i.e., being able to avoid fraud and corruption and unwilling to take advantage of other individuals or of society at large) and with the HIT sub-dimension Opposition-Defiance (i.e., being disrespectful for rules, laws, or authorities) showed convincing support for the construct validity of the SJT. No significant differences between correlations in the ethnic majority and minority group were found.

A first remarkable finding regarding Hypothesis 1 was the rather low correlation between the SJT and the dimension integrity as measured during the employment interview and comprised in the final employment recommendation. Although this finding was contrary to our expectation, two possible explanations may be given. One explanation is the difference in definitions of integrity used for the development of the SJT and the development of the employment interview (and thus also as used in the final recommendation). As can be seen in Appendix D, integrity as measured during the employment interview is defined in a quite broad and abstract way. Integrity as measured by the SJT, on the other hand, is defined very specifically, as very specific police integrity violations are used in the

scenarios. This lack of concordance in specificity and scope of the definitions may have caused the low correlation between the SJT and integrity as measured during the employment interview and between the SJT and integrity as comprised in the final recommendation. A second possible explanation for the low correlations is the small variance in scores on integrity as measured during the employment interview and on integrity as comprised in the final recommendation (see Table 4). This low variance probably suppressed the correlations. Therefore, the observed correlations between SJT and employment interview and SJT and final recommendation may be underestimates of the true correlations. Interviewers during the employment interview spend about three minutes collecting applicant information on integrity, which obviously is very short. Concerning integrity, interviewers were perhaps not very well equipped to properly differentiate between applicants, resulting in largely the same score for everyone.

A second remarkable finding pertaining to Hypothesis 1 was the larger than expected correlation between the SJT and the AC- and final-recommendation factor Agency. Agency consists of dimensions such as Flexibility and Decisiveness. As these dimensions are non-integrity related, the factor Agency was expected to have a low correlation with the SJT score. However, it is likely that to perform well on a SJT, i.e., to make the correct decisions with regard to the effectiveness of various responses in different situations, some decisiveness and flexibility is necessary. This might explain the higher than expected correlation between the SJT and Agency.

In accordance with the second hypothesis, we found a very small relationship between the SJT and cognitive ability. Regarding integrity tests, Ones and Viswesvaran (1998) showed that they have negligible correlations with cognitive ability. Since integrity was the intended SJT construct, we expected a small correlation between the SJT score and scores on the cognitive ability test. This was what we found, providing more evidence for the construct validity of the present SJT.

Finally, the results showed a score difference between the ethnic majority and minority group of .38 *SD*. This score difference was smaller than the score difference that was found on the cognitive ability test, which supported our third hypothesis. However, on integrity tests, Ones and Viswesvaran (1998) found even lower score differences (below .15 *SD*). The present finding of a .38 *SD* difference is identical to SJT findings by Nguyen et al. (2005) and O'Connell et al. (2007).

A small score difference could, on the one hand, be expected, because a video-based SJT was used, which tends to entail smaller score differences between ethnic groups than on paper-and-pencil SJTs (Chan & Schmitt, 1997). Lievens and Sackett (2006) argued that the difference between paper-and-pencil SJTs and video-based SJTs could be attributed to the greater reading component inherent to written SJTs. On the other hand, the effect of the video format on the score difference might be canceled out, because a knowledge instruction (i.e., "rate the effectiveness of each response") was used, which tends to bring about larger score differences than behavioral tendency instructions (i.e., "what would you most/least likely do?"; Nguyen & McDaniel, 2003). All in all, a small score difference between ethnic groups was expected, as score differences on integrity tests were found to be small (Ones & Viswesvaran, 1998). The score difference that was found in the present study was somewhat larger than the difference found by Ones and Viswesvaran (1998), although in size it is still considered as small. The somewhat larger score difference on the SJT than on other integrity tests, may be related to language skills. These skills are less important for the video-based SJT than the paper-and-pencil SJT but perhaps more important for the video-based SJT, with its complex verbal scenarios, than for an integrity test, with single sentence items.

Limitations

Our study had some limitations. First, the small sample size of ethnic minority applicants with regard to the in-depth Honesty-Humility interview and the HIT-questionnaire resulted in small power. With regard to the ethnic minority group, a larger sample size would have allowed stronger conclusions. Also, a larger sample size of ethnic minorities would allow a further differentiation within the ethnic minority group. De Meijer et al. (2006) showed that large differences on selection measures exist between ethnic minority groups, which might be explained by differences in history and culture between the ethnic groups. Investigating these ethnic minority groups separately, may result in more useful information compared to merely contrasting the ethnic majority to minority group and not taking into account potential differences between ethnic groups.

Second, we did not have criterion data at our disposal to investigate the criterion-related validity of the present SJT. Although the construct-validity results are promising, we do not know whether the present SJT is able to predict job performance, workplace (dis)honesty, theft, fraud, etc.. Since little is known about SJTs measuring a single construct, in general, and their criterion-related validity, specifically, future research should be focused on these types of SJTs and their predictive power. Furthermore, SJTs intended to

measure a single construct should be developed in different companies, in different settings, and on different job levels to be able to properly generalize the findings in the present study.

Conclusion

Contrary to the previous viewpoint of the construct-heterogeneity of SJTs, we argue that a construct-driven approach may be fruitful in the development of SJTs measuring one single construct. In a field study conducted in a multi-ethnical setting during Dutch police officer selection, we examined the construct validity of a video-based situational judgment test (SJT) measuring integrity. We investigated 1) the convergent and discriminant validity of the SJT, 2) correlations between the SJT score and scores on a cognitive ability test, and 3) the SJT score difference between the ethnic majority and the ethnic minority group. First, we found support for the construct validity of the Integrity-SJT. Second, we found a very small relationship between the SJT and the cognitive ability test. Finally, the results showed a SJT score difference of .38 *SD*, in favor of the ethnic majority group, which is in line with previous research findings on SJTs.

Chapter 7

Summary and discussion

The current dissertation presents five empirical studies investigating ethnic group differences on personnel selection measures used for the selection of Dutch police officers. These measures are a cognitive ability test, a Big Five personality questionnaire, an assessment center (AC), an employment interview, a final employment recommendation, and a situational judgment test (SJT). From the existing literature, little is known about why ethnic group differences in personnel selection exist. In the present dissertation the following potentially explanatory factors are distinguished: 1) applicant-related differences, 2) assessor-related differences, and 3) method-related factors (cf. Klimoski & Donahue, 2001). The largest ethnic minority groups in The Netherlands are examined, i.e., Dutch Antilleans, Moroccans, Surinamese, and Turks, which are, as a group or separately, compared to the Dutch majority group. Below, first, a succinct summary of the main research findings is given. This section will be followed by a more in-depth discussion of these findings.

7.1 Summary of Main Findings

In the introductory chapter, six research questions (RQs) were raised that directed the studies presented in this dissertation. These questions relate to the relative extent to which applicant (RQ1 and RQ2), assessor (RQ3 and RQ4), and selection-method factors (RQ5 and RQ6) are able to explain existing differences between ethnic groups. Guided by these six questions, the main findings of the five empirical studies are outlined below, based on a large-scale dataset (more than 13,000 applicants) from the Dutch police officer selection procedure.

Research Question 1: Do Score Differences on Selection Measures between Ethnic Groups Exist?
As a starting point for further research into possible explanatory factors for ethnic group differences in personnel selection, it was necessary to, firstly, determine whether ethnic differences in scores on selection measures exist. In the first part of **Chapter 2**, score differences on the various selection measures (i.e., a cognitive ability test, a personality questionnaire, an AC, an employment interview, and a final employment recommendation) were compared with each other. To this end, the four largest ethnic minority groups in The Netherlands each were compared to the Dutch majority group separately. In this comparison, first-generation minority groups were distinguished from second-generation minority groups.

Significant score differences between the ethnic majority group and first-generation ethnic minority groups to the advantage of the majority group existed on all selection tools. These differences were consistent with the findings from the literature. However, score differences between the majority group and the second-generation minority groups were much smaller, although the majority group still scored higher than the second-generation minority groups. The only exception to these score differences was the personality dimension Conscientiousness, on which minority groups systematically scored higher than the majority group. Score differences between the ethnic majority group and the ethnic minority groups were largest on the cognitive ability test and smallest on the personality questionnaire. Altogether, score differences on the personality questionnaire were rather unsystematic with sometimes the majority group and sometimes the minority group scoring higher. Score differences between the first-generation minority group and the second-generation minority group were the largest for the Antillean group, and the smallest for the Turkish group. Turkish minority applicants scored somewhat lower than the other ethnic

minority groups on all selection measures.

Research Question 2: Which Applicant Factors Play a Role in Ethnic Score Differences?
Investigating potentially explanatory applicant factors regarding the selection measure score differences, it was asked 1) whether applicant demographics (i.e., language proficiency, education, and ethnicity) were able to explain score differences on selection measures between ethnic groups, and 2) whether ethnic minority applicants whose demographics (i.e., language proficiency and education) were either very positive or very negative were evaluated more extremely (i.e., positive or negative, respectively) than ethnic majority applicants with the same very positive or negative demographics. In the second part of **Chapter 2**, the explanatory power was investigated of the applicant factors Dutch language-proficiency, education, and ethnicity on score differences on the so-called *objective measures* (i.e., the cognitive ability test and the personality questionnaire). Furthermore, regarding the so-called *subjective measures* (i.e., the AC, the employment interview, and the final employment recommendation), assumed-characteristics theory (Locksley, Borgida, Brekke, & Hepburn, 1980; Locksley, Hepburn, & Ortiz, 1982a; 1982b) and complexity-extremity theory (Linville, 1982; Linville & Jones, 1980) were tested.

Score differences on the objective measures were explained mostly by the applicant factor Dutch language-proficiency. Both education and ethnicity explained only small proportions of the score variance. Supportive of assumed-characteristics theory, results showed that the applicant factors Dutch-language proficiency and education explained more of the variance in score differences on subjective measures than did the applicant factor ethnicity. In addition, this finding was more outspoken when assessors had knowledge about the applicants' language-proficiency, education, and ethnicity compared to when they did not. Thus, having knowledge of someone's demographic characteristics diminishes the influence of ethnicity per se, as assumed-characteristics theory would predict. In contrast to what complexity-extremity theory would predict, ethnic minority applicants whose Dutch language-proficiency and education were either excellent or very low were not evaluated more extremely (i.e., positive or negative, respectively) than ethnic majority applicants with the same demographics. Rather, the general pattern was that the ethnic minority groups scored somewhat lower on the subjective measures.

Related to the effect of assessor factors on differences between ethnic groups of applicants, two research questions were formulated that guided the studies presented in Chapters 3 and 4. As assessor influence was the focus, these two

studies concentrated on subjective measures solely.

Research Question 3: Does the Judgment Process Differ when Assessing Different Ethnic Groups?

The first question concerning assessor factors (RQ3) asked whether assessors integrate information into a final employment recommendation differently when judging an ethnic majority applicant than when judging an ethnic minority applicant. To this end, in **Chapter 3**, a judgment-analysis study investigated assessors' judgment processes evaluating ethnic minority applicants and their judgment processes in evaluating ethnic majority applicants. The sample only contained assessors belonging to the ethnic majority group. The term 'judgment process' is used to describe the process of giving weights to sources of information (i.e., scores on an AC, an employment interview, and a personality questionnaire) when combining these into a final employment recommendation. Thus, the focus is not on subgroup score differences on selection measures, but on differences in weights when combining information from various selection measures into a final advice.

Results showed that the assessors used more irrelevant sources of information and based their decision to a lesser extent on their own ratings than on ratings of others when judging ethnic minority applicants compared to ethnic majority applicants. Probably because of less experience or more uncertainty when making a judgment about ethnic minority applicants, assessors used more and also more irrelevant information in their judgment process. They were also hesitant to use their own ratings when assessing ethnic minority applicants and incorporated information from others more in their decisions about ethnic minority applicants compared to majority applicants.

Research Question 4: Which Similarity Issues between Applicants and Assessors Play a Role in Score Differences?

A second question regarding the influence of assessor factors on differences between ethnic groups (RQ4) was directed at demographic and perceived similarity between assessor and applicant and at whether such similarity explains score differences between the ethnic majority group and ethnic minority groups. In **Chapter 4**, using multilevel analyses, the effects were investigated of actual demographic, i.e., ethnic, similarity between assessor and applicant and of perceived similarity of applicants by assessors on score differences on subjective measures. Assessors came from both the ethnic majority and the ethnic minority group.

From Social Identity Theory (Tajfel, 1982; Turner, 1987) and perceived intergroup similarity, it may be expected that both demographic and perceived similarity will lead to higher evaluations. However, previous research (e.g., McFarland, Ryan, Sacco, & Kriska, 2004; Strauss, Barrick, & Connerley, 2001) had found mixed results. Most studies did not use multilevel analysis, while for some studies it would have been the most appropriate analysis technique given their nested data structure. One reported study by Sacco, Scheu, Ryan, and Schmitt (2003) examined demographic similarity and analyzed their data with multilevel analyses. They found that demographic similarity between applicant and interviewer did not explain interview score differences between ethnic groups. The question, therefore, could be raised whether the same holds for other subjective measures and for perceived similarity. The study reported in Chapter 4 showed that neither demographic nor perceived similarity was able to explain score differences between the ethnic majority and the four ethnic minority groups on any of the subjective measures. Therefore, no evidence was found for (dis)similarity differentially affecting evaluations of ethnically diverse applicants during personnel selection.

The final two research questions focused on whether method factors, i.e., the psychological measures used, were able to explain differences between the ethnic majority group and ethnic minority groups.

Research Question 5: What is the Predictive Validity of Cognitive and Non-Cognitive Measures?

The first question regarding method factors (RQ5) explored whether the predictive validity of the non-cognitive ability tests was comparable to the predictive validity of the cognitive ability test for both the ethnic majority and the ethnic minority group. As was the case for RQ3, the research guided by RQ5 took a somewhat different approach than the direct investigation of potential effects on score differences. In **Chapter 5**, the differential validity of various selection measures was examined. A distinction was made between the cognitive ability test and non-cognitive ability tests (i.e., the personality questionnaire, the AC, the employment interview, and the final employment recommendation).

In contrast to the mostly U.S.-dominant literature on differential prediction until now, results indicated differential prediction on the cognitive ability test as well as on several non-cognitive ability tests, namely the AC, the employment interview, and the final employment recommendation. The cognitive ability test appeared to be more predictive of training success for ethnic minority trainees than for ethnic majority trainees. Yet, the AC, the

interview, and the final recommendation were more predictive for ethnic majority trainees than for minority trainees. The personality questionnaire showed very little predictive power for both ethnic groups.

Research Question 6: Can Situational Judgment Tests Measure Integrity across Ethnic Groups?
The second question concerning method factors (RQ6) studied to what extent a situational judgment test (SJT) developed to measure Integrity, indeed measured the same construct for the ethnic majority group and the ethnic minority group. In the final empirical chapter (**Chapter 6**), a study was described focusing on the construct-driven development of an Integrity-SJT, which was based on video scenarios.

Although previous research had found that SJTs are often construct-heterogeneous, the study reported in Chapter 6 found support for the construct validity of the Integrity-SJT for both ethnic groups. Furthermore, the results showed that the SJT score difference was substantially smaller than the score difference that is generally found on cognitive ability tests. The SJT, thus, appeared to be a useful measure of Integrity in a multi-cultural setting.

7.2 Discussion, Practical Implications, and Future Research

The main purpose of the present dissertation was to increase our knowledge about potential factors explaining differences between ethnic groups in personnel selection. The importance of determining explanatory factors of ethnic subgroup differences in selection is twofold. First, labor-force and employment numbers exhibit a relatively low presence of ethnic minority group members in the Dutch working population compared to the ethnic majority group (CBS, January 1, 2007). Investigating specific factors related to ethnic differences may provide useful information regarding the role of personnel selection processes on the employment opportunities of specific ethnic groups. Second, because personnel selection has a substantial impact on subsequent employment, it is important to increase our understanding of the factors that may influence personnel selection among ethnic majority and minority applicants. That is, to what extent are applicant-related differences, assessor-related differences, and selection-method factors important in explaining differences in selection and, thus, employment opportunities between ethnic groups? In the following sections, the research findings will be discussed in this light. Furthermore, practical implications and directions for future research will be outlined.

Applicant Factors

Possible explanatory ethnicity-related applicant characteristics of score differences between ethnic groups were investigated. Score differences between the ethnic majority group and ethnic minority groups appeared to be much smaller for second-generation minority groups than for first-generation minority groups, with the applicant factor Dutch language-proficiency explaining most of the variance in score differences between the ethnic majority and minority group (for both first- and second-generation minority applicants). Therefore, it is quite plausible that the score differences between first- and second-generation minority applicants also can be explained by an improved Dutch language-proficiency from one generation to the next. Research by Bleichrodt and Van den Berg (1995) supports these findings. They found that first-generation ethnic minorities who moved to The Netherlands before the age of seven (before starting their primary education) scored significantly higher on cognitive ability tests than first-generation ethnic minorities who moved to The Netherlands after the age of seven. Passing through a large part of the Dutch educational system most probably had improved the Dutch language-proficiency of the first group of first-generation minorities. It can also be argued that the skills (e.g., language skills, math skills) learned during their Dutch education as well as the cultural aspects of The Netherlands had improved their test results. This may explain why, from one generation to the next, score differences diminished on all measures, that is, not only on the cognitive ability test.

Other ethnicity-related applicant factors, not examined in this dissertation, which may have influence on score differences are socio-economic status (Hofman, 1993; Van der Velden, 1991) or the home interference with (school)work (e.g., Demerouti, Bakker, & Schaufeli, 2005; Van Emmerik & Jawahar, 2006). When ethnicity-related applicant factors, but not ethnicity per se, appear to influence ethnic score differences and when these score differences diminish from one generation to the next, the idea in the U.S. of the Bell Curve (Herrnstein & Murray, 1994) that argues that Black-White score differences on *g*-loaded tests can at least to some extent be attributed to genetic differences, becomes more and more groundless.

Assessor Factors

Concerning assessor factors, Chapters 3 and 4 focused on the subjective measures. A judgment-analysis study showed that assessors, who all were from the ethnic majority group, used a larger number of irrelevant cues for the judgment of ethnic minority applicants than for the judgment of ethnic majority applicants. Furthermore, when judging ethnic minority applicants, assessors based their decision to a lesser extent on their own ratings than on

ratings of others. It is argued in Chapter 3 that the difference in judgment process is caused by less experience and more insecurity of assessors when it comes to assessing ethnic minority applicants in personnel selection.

Some assessors, however, remained consistent in the use of information across ethnic groups. For these assessors, the ethnicity of the applicant does not seem to make a difference in information processing during personnel selection. It may be speculated, as no direct empirical study has been conducted, that those assessors who do not differentiate between ethnic majority and minority applicants in their judgment processes are better assessors. Funder (1995), for instance, argued that a 'good judge' was characterized by having knowledge about correct cues (or information sources). Systematic cue use across ethnic groups may form evidence for correct cue use, as unsystematic cue use across ethnic groups implies that at least for one ethnic group the judgment process is less correct.

When equal job opportunities for ethnic majority and minority applicants are aimed for, the finding of differences in judgment processes of assessors in selection settings may have important implications for practice. Pursuing equal opportunities in personnel selection, more exchange of knowledge about assessment in a multi-cultural setting among assessors may be necessary. Furthermore, the integration of different sources of information into a final recommendation could be realized in a more standardized manner, for instance, statistically instead of clinically (Grove, Zald, Lebow, Snitz, & Nelson, 2000). As a consequence, differences across ethnic groups in the decision-making process to hire or reject applicants may diminish.

Judgment-analysis studies focus on differences in evaluations by different assessors. The present judgment-analysis research in Chapter 3 demonstrated the existence of individual differences in the way assessors weigh and combine information about applicants. Other important aspects related to applicant ethnicity during personnel selection are demographic similarity and perceived similarity between assessors and applicants. The effects of attitudes towards applicants from different ethnic groups on scores given by the assessors were studied in Chapter 4. To this end, demographic similarity between assessor and applicant, and perceived similarity of the applicant by the assessor were examined. Results from multilevel analysis showed that neither demographic nor perceived similarity was able to explain score differences between the ethnic majority and the four ethnic minority groups on the subjective measures. These findings are in line with previous findings by Sacco et al. (2003), using the same multilevel analysis technique.

Explanations for the lack of effects of demographic and perceived similarity on given ratings are that during personnel selection raters have a strong motive to be accurate and that well-trained raters have learned to focus on a largely structured task and, therefore, will be less influenced by aspects of (dis)similarity. For practitioners, these findings, fortunately, alleviate concerns that discrimination of ethnic minority groups due to (dis)similarity may occur during personnel selection.

Selection-Method Factors

The final two empirical studies are reported in Chapters 5 and 6 and investigated the influence of selection-method factors on differences between ethnic groups. Pertaining to the issue of differential predictive validity, in Chapter 5, the predictive validity of various cognitive and non-cognitive ability selection measures was examined. Contrary to previous research findings, which were mostly based on U.S. samples and reported no differential prediction, evidence was found for differences in validity for the ethnic majority and minority group for both cognitive and non-cognitive ability measures. For the ethnic minority group, training performance was mainly predicted by the cognitive ability test. For the ethnic majority group, in contrast, the cognitive ability test showed very little predictive power. Non-cognitive ability variables showed much more predictive validity in this group.

One possible explanation for the relatively low validities of cognitive ability tests that is found in previous research on police work, lies in the potential role of non-cognitive factors in the determination of performance as stated by Hirsh et al. (1986). Interestingly, however, in the present study this explanation only pertains to the ethnic majority group, for which various non-cognitive ability factors were more predictive for training performance than cognitive ability. For the minority group, however, the cognitive ability test was most predictive, in particular the verbal cognitive ability subtests, i.e., Verbal Comprehension and Word Fluency. The non-cognitive ability tests showed very little predictive power for this latter group.

The differential validity might be caused by ethnic bias of ethnic majority supervisors' subjective evaluations (Te Nijenhuis & Van der Flier, 2000), even though evaluations of trainees during the Dutch police training were structured according to evaluation forms. Apparently, for ethnic majority trainees, relatively more attention seems to be given to the non-cognitive ability aspects of performance, such as social skills, decisiveness, and authority. While, for ethnic minority trainees, relatively more attention seems to be given to the verbal cognitive ability aspects of performance. The question to be looked into remains whether supervisors' evaluations of ethnic

minority trainees are predominantly susceptible to these basic language skills to the extent that these skills will overshadow other important non-cognitive factors, such as social skills and decisiveness. To better understand potential supervisors' susceptibility to ethnic bias, research using ethnic majority as well as ethnic minority supervisors should receive more attention in the future.

A type of bias that has recently received much attention is what Jencks (1998) labels 'selection system bias'. Selection system bias occurs when there are larger subgroup differences on a gating mechanism than on the behavior being predicting by that gating mechanism (Crosby, Iyer, & Sincharoen, 2006). In the study presented in Chapter 5, the score differences on the selection measures were much larger than on the training criteria, which may be evidence for this selection system bias. In an overview of recent meta-analytical findings, Gamliel and Cahan (2007) argued that a possible explanation for this differential gap is that selection measures are often standardized while typical measures used as work-related criteria are non-standardized subjective evaluations. In a similar vein, Roth, Huffcutt, and Bobko (2003) meta-analytically showed that ethnic score differences were larger on objective criteria than on subjective criteria. It is common practice to uncover the psychometric properties of predictors that are used in high-stakes situations. However, the findings of Gamliel and Cahan (2007), Te Nijenhuis and Van der Flier (2000), and Chapter 5 imply that future research should be directed at determining the psychometric properties of criteria as well.

Finally, in Chapter 6, the construct validity was investigated of a newly developed video-based situational judgment test (SJT) aimed to validly measure Integrity for the ethnic majority as well as the ethnic minority group. Because of its characteristics, the SJT as a selection method has become increasingly popular in personnel selection and in the research literature during the last two decades (e.g., Chan & Schmitt, 1997, 2005; McDaniel, Hartman, Whetzel, & Grubb, 2007). These are high criterion-related validity (McDaniel et al., 2007), new technology that has made the development of video-based SJTs possible, and, finally, little adverse impact against ethnic minority groups (e.g., Weekley & Jones, 1997, 1999). Also, higher realism of video-based SJTs was found to lead to more reliable respondent reactions (Chan & Schmitt, 1997; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000). Because of these advantages, a video-based SJT was developed in a multi-ethnical setting in The Netherlands.

Despite the qualities of SJTs, one critical issue is the often-found construct-heterogeneity of SJTs and the difficulty of developing a SJT that measures

one specific construct. In the study presented in Chapter 6, it was investigated whether it would be possible to develop a SJT measuring one specific construct, in this case Integrity, which at the same time shows little adverse impact. Support, indeed, was found for the construct validity of the Integrity-SJT in both the ethnic majority and minority group. Furthermore, ethnic score differences on the Integrity-SJT were substantially smaller than on the cognitive ability test. In recent years, a lot of research had been conducted on the topic of the SJT. As this type of test appears to be promising for personnel-selection practices in a multi-ethnical setting, future research should focus more strongly on the possible construct-homogeneity of SJTs intended to measure other constructs than Integrity.

7.3 Conclusion

In conclusion, the research reported in the present dissertation has highlighted several issues. One important issue is language as Dutch language-proficiency of applicants explained a substantial part of the score differences between the ethnic majority group and ethnic minority groups. Interestingly, assessor-applicant (dis)similarity did not differentially affect evaluations of ethnically diverse applicants. This finding alleviates concerns that discrimination of ethnic minority groups due to (dis)similarity may occur during personnel selection. However, a difference was found in the decision-making process of ethnic majority assessors judging ethnic minority applicants compared to ethnic majority applicants. This finding indicates that assessors are, in some way, affected by the ethnicity of applicants. Gaining experience in assessing ethnic minority applicants, exchanging knowledge about assessment in a multi-cultural setting among assessors, or perhaps further standardizing the selection process should diminish differential effects. Furthermore, selection measures, both cognitive and non-cognitive, appear to differentially predict training performance of ethnic majority and minority trainees. A possible explanation of this differential effect may lay in the subjective evaluations of supervisors during training. Finally, scores on a newly developed situational judgment test (SJT) turned out to show substantially smaller ethnic group differences than generally are found on the cognitive ability test. These findings yield practical guidelines for personnel selection in a multi-cultural setting, such as further standardization of the decision-making process to hire or reject applicants and diminishing the influence of language skills of applicants by means of SJTs. More research is needed to further improve our understanding of personnel selection, specifically, and job opportunities, in general, in a multi-cultural setting.

Nederlandse samenvatting

Het bekende cliché bij de Nederlandse politie luidt: Marokkaanse verdachten bekennen nooit, zijn niet aanspreekbaar op hun gedrag en reageren per definitie fel (Politieacademie, 2007). Toch kun je een Marokkaan best aan het praten krijgen, aldus een agent van politiekorps Utrecht. "Maar dan moet je zijn culturele normen en waarden kennen. Belangrijk is dat je in contact probeert te komen. Ga de dialoog aan, niet de confrontatie. Vanuit die persoonlijke band kun je gevoelige zaken boven tafel krijgen." (Politieacademie, 2007, p. 12). Diezelfde politieagent noemt als één van de voorbeelden waarbij zijn Marokkaanse achtergrond hem voordelen biedt in het politiewerk: "Ik ben een aanspreekpunt binnen de politie. Voor de Marokkaanse gemeenschap zelf, maar ook voor collega's. We hebben te maken met jongeren die opgegroeid zijn in twee culturen en vanuit daar een eigen mengcultuur hebben ontwikkeld. Ik ken die en kan collega's erover adviseren." (Politieacademie, 2007, p. 12).

Bovenstaand voorbeeld toont de voordelen van het aantrekken van allochtone werknemers door de Nederlandse politie. Allochtonen worden hier gedefinieerd als diegenen die in het buitenland geboren zijn of van wie één van de ouders in het buitenland geboren is. Autochtonen zijn diegenen die in Nederland geboren zijn. Daarbij zijn ook de ouders in Nederland geboren (CBS, 1 januari, 2007). Een allochtone politieagent heeft kennis van de taal, cultuur en religie van de allochtone groep waartoe hij of zij behoort. Daarnaast blijkt het voor een allochtone politieagent gemakkelijker te zijn om allochtone burgers aan te spreken op ongewenst gedrag dan voor een autochtone agent en heeft een allochtone politieagent vaak een voorbeeldfunctie voor de allochtone jeugd (Broekhuizen, Raven, & Driessen, 2007).

Ondanks de voordelen van het in dienst nemen van allochtone medewerkers, is het personeelsbestand van de Nederlandse politie geen afspiegeling van de Nederlandse samenleving (LECD, 2006). De Nederlandse politie is hierin echter geen uitzondering. De allochtone bevolking in Nederland blijkt in het algemeen ondervertegenwoordigd in de beroepsbevolking (CBS, 1 januari, 2007). Dit kan verschillende oorzaken hebben, zoals wervingscampagnes die allochtonen niet bereiken, personeelsselectie die nadelig is voor allochtone kandidaten of carrièreperspectieven die minder rooskleurig zijn voor de allochtone groep dan voor de autochtone groep. Dit proefschrift richt zich op

personeelsselectie. In de bestaande literatuur wordt op dit moment nog weinig naar mogelijke verklaringen gezocht voor het bestaan van etnische groepsverschillen op het gebied van personeelsselectie. De volgende mogelijke verklarende factoren worden onderscheiden in dit onderzoek naar verschillen tussen de autochtone en de allochtone groep: 1) individuele kandidaat-gerelateerde verschillen, 2) beoordelaargerelateerde verschillen en 3) selectie-instrument-gerelateerde verschillen (cf. Klimoski & Donahue, 2001). Het onderzoek dat gepresenteerd wordt in dit proefschrift werd uitgevoerd bij de Nederlandse politie, waarbij gegevens van ruim dertienduizend kandidaten gebruikt werden. De verschillende allochtone groepen zijn de grootste allochtone groepen in Nederland, namelijk Antillianen, Marokkanen, Surinamers en Turken.

De in dit proefschrift beschreven empirische studies zijn gericht op verschillende factoren die mogelijk van invloed kunnen zijn op verschillen die bestaan tussen de autochtone groep en de allochtone groepen. De studie in hoofdstuk 2 heeft betrekking op het effect van kandidaat-kenmerken op verschillen tussen etnische groepen en de studies in de hoofdstukken 3 en 4 op beoordelaarkenmerken. De studies in de hoofdstukken 5 en 6, tenslotte, zijn uitgevoerd om het effect van kenmerken van verschillende selectie-instrumenten nader te onderzoeken.

Overzicht van Empirische Bevindingen

Als start voor het onderzoek naar mogelijke verklarende factoren van etnische verschillen bij personeelsselectie, werd bekeken hoe groot de verschillen in scores op selectie-instrumenten daadwerkelijk zijn. In het eerste gedeelte van **hoofdstuk 2** werden daartoe de scoreverschillen op verschillende selectie-instrumenten (dat wil zeggen een cognitieve capaciteitentest, een persoonlijkheidsvragenlijst, een assessment center [AC], een selectie-interview en een selectie-eindadvies) berekend en met elkaar vergeleken. Daarnaast werden de vier grootste allochtone groepen in Nederland – namelijk Antillianen, Marokkanen, Surinamers en Turken – elk afzonderlijk vergeleken met de autochtone groep. Hierbij werd tot slot onderscheid gemaakt tussen eerste-generatie allochtone groepen (dat wil zeggen zelf in het buitenland geboren) en tweede-generatie allochtone groepen (dat wil zeggen zelf in Nederland geboren, maar minstens één van de ouders in het buitenland geboren).

Op alle selectie-instrumenten en consistent met bevindingen uit de literatuur werden scoreverschillen gevonden tussen de autochtone groep en eerste-generatie allochtone groepen, in het voordeel van de autochtone groep. De scoreverschillen tussen de autochtone groep en tweede-generatie allochtone

groepen waren echter veel kleiner. De autochtone groep scoorde wel hoger dan de tweede-generatie allochtone groepen. De enige uitzondering hierop betrof de persoonlijkheidsdimensie Consciëntieusheid, waarop de allochtone groepen systematisch hoger scoorden dan de autochtone groep. Scoreverschillen tussen de autochtone groep en allochtone groepen waren het grootst op de cognitieve capaciteitentest en het kleinst op de persoonlijkheidsvragenlijst. Over het algemeen waren de scoreverschillen op de persoonlijkheidsdimensies vrij onsystematisch; soms scoorde de autochtone groep hoger en soms de allochtone groepen. Scoreverschillen tussen de eerste-generatie allochtone groepen en de tweede-generatie allochtone groepen waren het grootst voor de Antilliaanse groep en het kleinst voor de Turkse groep. Turkse kandidaten scoorden iets lager dan andere allochtone groepen op alle selectie-instrumenten.

Toen bleek dat op alle selectie-instrumenten verschillen bestonden tussen autochtone en allochtone kandidaten die bovendien vrij substantieel waren, werd gezocht naar mogelijke verklaringen. Daartoe richtte het onderzoek zich op factoren die mogelijk van invloed kunnen zijn op deze verschillen. Met betrekking tot kandidaat-kenmerken werd antwoorden op de volgende vragen gezocht: 1) in hoeverre verklaren taalkennis, opleiding en etniciteit verschillen in scores tussen autochtone en allochtone kandidaten op selectie-instrumenten? en 2) worden allochtone kandidaten met zeer positieve of negatieve kenmerken (in termen van taalkennis en opleiding) extremer (dat wil zeggen respectievelijk positiever of negatiever) beoordeeld dan autochtone kandidaten met dezelfde kenmerken? In het tweede deel van **hoofdstuk 2** werd de verklarende kracht onderzocht van de kandidaat-kenmerken Nederlandse taalkennis, opleiding en etniciteit op scoreverschillen op de zogenaamde *objectieve instrumenten* (dat wil zeggen de cognitieve capaciteitentest en de persoonlijkheidsvragenlijst). Daarnaast werden, met betrekking tot de zogenaamde *subjectieve instrumenten* (dat wil zeggen het AC, het selectie-interview en het selectie-eindadvies), de veronderstelde-kenmerkentheorie (*assumed-characteristics theory*; Locksley, Borgida, Brekke & Hepburn, 1980; Locksley, Hepburn & Ortiz, 1982a; 1982b) en de complexiteit-extremiteittheorie (*complexity-extremity theory*; Linville, 1982; Linville & Jones, 1980) op het gebied van personeelsselectie getest.

De resultaten toonden dat scoreverschillen op de objectieve instrumenten voornamelijk verklaard werden door het kandidaat-kenmerk Nederlandse taalkennis. Zowel opleiding als etniciteit verklaarden slechts een klein deel van de variantie. In overeenstemming met de veronderstelde-kenmerkentheorie toonden de resultaten met betrekking tot de subjectieve instrumenten dat de kandidaat-kenmerken Nederlandse taalkennis en opleiding meer variantie in

scoreverschillen verklaarden dan het kandidaat-kenmerk etniciteit. Bovendien bleek deze bevinding meer uitgesproken te zijn wanneer beoordelaars op de hoogte waren van de taalkennis, opleiding en etniciteit van de kandidaat dan wanneer zij daarvan niet op de hoogte waren. Geconcludeerd kan worden dat kennis van iemands demografische kenmerken het effect van etniciteit op zich vermindert. In tegenstelling tot wat de complexiteit-extremiteittheorie voorspelt, werden allochtone kandidaten van wie de Nederlandse taalkennis en opleiding of heel hoog of heel laag waren niet extremer (dat wil zeggen respectievelijk positiever of negatiever) geëvalueerd dan autochtone kandidaten met dezelfde kenmerken. Voor de allochtone groep leek een algemene tendens te bestaan om enigszins lager beoordeeld te worden op de subjectieve instrumenten.

Omdat in de hoofdstukken 3 en 4 het effect van potentiële beoordelaarkenmerken op etnische groepsverschillen werd onderzocht, beperkten de studies uit deze hoofdstukken zich tot de subjectieve instrumenten. Antwoord werd gezocht op de volgende vragen: 1) in hoeverre integreren beoordelaars selectie-informatie in een selectie-eindadvies verschillend wanneer zij een autochtone of een allochtone kandidaat beoordelen? en 2) in hoeverre verklaren demografische en waargenomen gelijkheid tussen beoordelaar en kandidaat scoreverschillen tussen autochtone en allochtone kandidaten? Met betrekking tot de eerste vraag werd in **hoofdstuk 3** een studie uitgevoerd waarbij gebruik werd gemaakt van beoordelingsanalyse (*judgment analysis*). Het beoordelingsproces van autochtone beoordelaars bij het beoordelen van autochtone of allochtone kandidaten werd onderzocht. Met de term 'beoordelingsproces' wordt het proces bedoeld van het toekennen van gewichten aan informatiebronnen (dat wil zeggen scores op een AC, een selectie-interview en een persoonlijkheidsvragenlijst) wanneer deze gecombineerd worden in een selectie-eindadvies.

Voor deze studie waren uitsluitend autochtone beoordelaars waren beschikbaar. De vergelijking tussen autochtone en allochtone kandidaten liet de volgende resultaten zien. De beoordelaars gebruikten meer irrelevante bronnen van informatie bij het beoordelen van allochtone kandidaten. Met een irrelevante informatiebron wordt hier een bron van informatie bedoeld die bij een eindadvies op een bepaalde dimensie niet relevant wordt geacht. Ook baseerden de beoordelaars hun beslissingen in mindere mate op hun eigen beoordelingen dan op beoordelingen van anderen wanneer zij allochtone kandidaten beoordeelden. Met eigen beoordelingen worden hier de interviewscores bedoeld, omdat de beoordelaar die het eindadvies van een bepaalde kandidaat vormt, ook het selectie-interview met deze kandidaat heeft

gevoerd. Beoordelingen van anderen zijn afkomstig van de persoonlijkheidsvragenlijst (zelfrapportage door de kandidaat) en het AC (beoordeling door een andere beoordelaar). De verklaring voor deze bevindingen werd gezocht in het feit dat de beoordelaars minder ervaring en daardoor meer onzekerheid hadden in het beoordelen van allochtone kandidaten. Zij twijfelen wellicht aan hun eigen beoordelingen bij allochtone kandidaten en gebruikten meer informatie afkomstig van anderen bij hun beslissingen over allochtone kandidaten vergeleken met autochtone kandidaten.

Met betrekking tot de tweede vraag over beoordelaarkenmerken, werd in **hoofdstuk 4** het effect onderzocht van demografische – hier: etnische – gelijkheid tussen beoordelaar en kandidaat en waargenomen gelijkheid van kandidaten door beoordelaars op de scoreverschillen op de subjectieve instrumenten. Zowel autochtone als allochtone beoordelaars deden mee aan dit onderzoek. Bij de studie in hoofdstuk 4 werd gebruik gemaakt van multi-levelanalyse.

Vanuit de sociale-identiteittheorie (*social identity theory*; Tajfel, 1982; Turner, 1987) en waargenomen-intergroepsgelijkheid (*perceived intergroup similarity*) werd verwacht dat zowel demografische als waargenomen gelijkheid zou leiden tot hogere beoordelingen. Echter, eerder onderzoek (bijvoorbeeld McFarland, Ryan, Sacco & Kriska, 2004; Strauss, Barrick & Connerley, 2001) vond geen eenduidig resultaat. Bovendien gebruikten de meeste studies geen multi-levelanalyse, terwijl dit voor sommige studies de meest geschikte analysetechniek zou zijn geweest, gegeven de geneste structuur van de data. Eén studie van Sacco, Scheu, Ryan en Schmitt (2003) onderzocht demografische gelijkheid en gebruikte wel multi-levelanalyse. Zij vonden dat demografische gelijkheid tussen kandidaat en interviewer geen effect had op scoreverschillen tussen etnische groepen bij het selectie-interview. De vraag rees of hetzelfde geldt voor andere subjectieve instrumenten en voor waargenomen gelijkheid. De studie in hoofdstuk 4 liet zien dat noch demografische noch waargenomen gelijkheid scoreverschillen op de subjectieve instrumenten verklaart tussen autochtone en allochtone kandidaten. Derhalve werd geen bewijs gevonden voor het differentiële effect van gelijkheid op beoordelingen van etnisch verschillende kandidaten tijdens personeelsselectie.

In de hoofdstukken 5 en 6 werd onderzocht of potentiële selectie-instrument-gerelateerde factoren, dat wil zeggen de psychologische instrumenten die zijn gebruikt, verschillen verklaren tussen de autochtone groep en allochtone groepen. De volgende vragen werden gesteld: 1) is de predictieve validiteit

van de cognitieve capaciteitentest vergelijkbaar met de predictieve validiteit van niet-cognitieve instrumenten voor zowel de autochtone als de allochtone groep? en 2) meet de situationele inzichttest (*situational judgment test* [SJT]), ontwikkeld om integriteit te meten, daadwerkelijk hetzelfde construct voor de autochtone en de allochtone groep? Met betrekking tot de eerste vraag werd in **hoofdstuk 5** de differentiële validiteit van verscheidene selectie-instrumenten onderzocht, namelijk de cognitieve capaciteitentest enerzijds en de persoonlijkheidsvragenlijst, het AC, het selectie-interview en het eindadvies anderzijds.

In tegenstelling tot de bestaande literatuur over differentiële predictie, die voornamelijk afkomstig is uit de V.S., bleken de resultaten in hoofdstuk 5 te wijzen in de richting van differentiële predictie voor zowel de cognitieve capaciteitentest als voor verscheidene niet-cognitieve instrumenten, namelijk het AC, het selectie-interview en voor het eindadvies. De cognitieve capaciteitentest bleek later opleidingssucces beter te voorspellen voor de allochtone groep dan voor de autochtone groep. Het AC, het interview en het eindadvies bleken daarentegen betere voorspellers te zijn voor de autochtone groep dan voor de allochtone groep. De persoonlijkheidsvragenlijst toonde erg weinig voorspellende kracht voor beide groepen.

Tot slot werd in **hoofdstuk 6** de constructgerichte ontwikkeling beschreven van een Integriteit-SJT die gebaseerd is op videoscenario's. Hoewel eerder onderzoek heeft aangetoond dat SJTs vaak constructheterogeen zijn, werd in de studie in hoofdstuk 6 steun gevonden voor de constructvaliditeit van de Integriteit-SJT in zowel de autochtone als de allochtone groep. Daarnaast toonden de resultaten dat het scoreverschil tussen de autochtone en allochtone groep op de SJT substantieel kleiner was dan het scoreverschil dat vaak gevonden wordt op de cognitieve capaciteitentest. Deze resultaten wijzen erop dat de SJT een zinvol instrument is om integriteit te meten in een multi-etnische context.

Conclusies

Tot besluit belichtte het onderzoek dat is gerapporteerd in dit proefschrift verschillende zaken, die belangrijk zijn voor selectie in een multi-etnische context. Een belangrijke kwestie bleek taal te zijn, omdat Nederlandse taalkennis een substantieel deel van de scoreverschillen tussen de autochtone en de allochtone groepen verklaarde. Interessant is voorts dat beoordelaar-kandidaatgelijkheid geen differentieel effect had op de beoordelingen van etnisch diverse kandidaten. Dit gold voor zowel autochtone als allochtone assessoren. Deze bevinding verlicht de zorg enigszins dat discriminatie plaatsvindt van allochtone kandidaten als gevolg van ongelijkheid tussen

beoordelaar en kandidaat. Er werd echter wel een verschil in beoordelingsproces gevonden tussen beoordelaars die autochtone kandidaten evalueerden en beoordelaars die allochtone kandidaten evalueerden. Dit resultaat laat zien dat sommige beoordelaars toch in zekere zin beïnvloed worden door de etniciteit van kandidaten. Het opdoen van ervaring bij het beoordelen van allochtone kandidaten, kennis uitwisselen tussen beoordelaars onderling over beoordelen in een multiculturele context en wellicht een verdere standaardisatie van het selectieproces zouden deze verschillen moeten verminderen. Verder bleken selectie-instrumenten, zowel cognitieve als niet-cognitieve, opleidingssucces verschillend te voorspellen voor de autochtone groep en voor de allochtone groep. Een mogelijke verklaring voor dit differentiële effect zou kunnen liggen in de verschillende subjectieve evaluaties van supervisoren die tijdens de opleiding de studenten beoordelen op hun prestaties. Tot slot bleken de scores op een nieuw ontwikkelde situationele inzichttest (*situational judgment test* [SJT]) substantieel kleinere groepsverschillen te vertonen dan vaak gevonden wordt op de cognitieve capaciteitentest.

Deze bevindingen resulteren in een aantal praktische richtlijnen voor personeelsselectie in een multiculturele context, zoals verdere standaardisatie van het besluitvormingsproces voor aannemen of afwijzen van kandidaten en het verminderen van de rol van de taalvaardigheid van kandidaten door bijvoorbeeld het gebruik van SJTs. Desalniettemin is verder onderzoek nodig om ons begrip te vergroten van de processen die plaatsvinden bij multi-etnische personeelsselectie in specifieke zin en van verschillende kansen van autochtone en allochtone groepen op de werkvloer in algemene zin.

References

Anderson, N., & Ones, D. S. (2003). The construct validity of three entry level personality inventories used in the UK: Cautionary findings from a multiple inventory investigation. *European Journal of Personality, 17*(1), 39-66.

Arbuckle, J. L. (2003). Amos 5.0 [Computer software]. Chicago: Smallwaters.

Arbuckle, J. L. (2005). Amos 6.0 [Computer software]. Chicago: SPSS Inc.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6*(3), 267-285.

Barr, S. H., & Hitt, M. A. (1986). A comparison of selection decision models in manager versus student samples. *Personnel Psychology, 39*(3), 599-617.

Barrick, M. R., Mount, M. K., & Judge, T. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1/2), 9-30.

Barriga, A. Q., & Gibbs, J. C. (1996). Measuring cognitive distortions in antisocial youth: Development and preliminary validation of the 'How-I-Think' questionnaire. *Aggressive Behavior, 22*(3), 333-343.

Barriga, A. Q., Gibbs, J. C., Potter, G. B., & Liau, A. K. (2001). *The 'How-I-Think' Questionnaire: Manual.* Champaign, IL: Research Press.

Bass, B. M., & Barrett, G. V. (1981). *People, work, and organizations. An introduction to industrial and organizational psychology* (2nd ed.). Boston: Allyn & Bacon.

Becker, T. E. (2005). Development and validation of a situational judgment test of employment integrity. *International Journal of Selection and Assessment, 13*(3), 225-232.

Bentler, P. M. (1990). Comparative fit indexes via structural models. *Psychological Bulletin, 107*(2), 238-246.

Berry, J. W. (1997). Immigration, acculturation, and adaptation. *Applied Psychology: An International Review, 46*(1), 5-68.

Bleichrodt, N., & Van den Berg, R. H. (1995). *Multiculturele capaciteitentest middelbaar niveau (MCT-m). Handleiding* [Multicultural capacity test medium level. Manual]. Amsterdam: Stichting NOA.

Bobko, P., Roth, P. L., & Bobko, P. (2001). Correcting the effect size of *d* for range restriction and unreliability. *Organizational Research Methods, 4*(1), 46-61.

Bollen, K. A. (1989). A new incremental fit index for general structural models. *Sociological Methods and Research, 17*(3), 303-316.

Bors, D. A., & Forrin, B. (1995). Age, speed of information processing, recall, and fluid intelligence. *Intelligence, 20*(3), 229-248.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment. The SJT view* (pp. 75-114). Amsterdam: North Holland.

Brehmer, B. (1988). The development of social judgment theory. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment. The SJT view* (pp. 13-40). Amsterdam: North Holland.

Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica, 87*(2/3), 137-154.

Broekhuizen, J., Raven, J., & Driessen, F. M. H. M. (2007). *Positie en expertise van de allochtone politiemedewerker* [Position and expertise of the ethnic minority police employee]. Utrecht, The Netherlands: Bureau Driessen.

Brunswick, E. (1952). *The conceptual framework of psychology*. Chicago, IL: University of Chicago Press.

Bureau InterCulturele Evaluatie (2000). *Intaketoets beroepsopleidingen op cd-rom.* [Intaketest for vocational education on cd-rom]. Lienden, The Netherlands.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. Amsterdam: North Holland.

Center for Multilevel Modeling (1997). MLwiN 1.10 [Computer software]. Institute of Education, London.

Central Bureau of Statistics (CBS; 2004). *Mens en maatschappij - onderwijs: Aandeel niet-westerse allochtonen in examenklassen voortgezet onderwijs, naar schoolgemeente of woongemeente.* [Men and society - education: Number of non-western ethnic minority high school students in final classes per school or area]. Voorburg/Heerlen, The Netherlands: CBS.

Central Bureau of Statistics (CBS). *Bevolking naar herkomstgroepering en generatie* [Population's ethnic background and generation]. (2007, January 1). Retrieved July 3, 2007, from http://www.cbs.nl

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perception. *Journal of Applied Psychology, 82*(1), 143-159.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*(3), 233-254.

Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, N. Anderson, & O. Voskuijil (Eds.), *Handbook of personnel selection* (pp. 219-246). Oxford: Blackwell.

Chattopadhyay, P., Tluchowska, M., & George, E. (2004). Indentifying the ingroup: A closer look at the influence of demographic dissimilarity on employee social identity. *Academy of Management Review, 29*(2), 180-202.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*(3), 410-417.

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

Coleman, L. M., Jussim, L., & Kelley, S. H. (1995). A study of stereotyping: Testing three models with a sample of blacks. *Journal of Black Psychology, 21*(4), 332-356.

Cortina, J. M., Doherty, M. L., Schmitt, N., Kaufman, G., & Smith, R. G. (1992). The "Big Five" personality factors in the IPI and MMPI: Predictors of police performance. *Personnel Psychology, 45*(1), 119-140.

Crosby, F. J., Iyer, A., & Sincharoen, S. (2006). Understanding affirmative action. *Annual Review of Psychology, 57*, 585-611.

Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology, 9*(1), 23-32.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95-106.

Dayan, K., Kasten, R., & Fox, S. (2002). Entry-level police candidate assessment center: An efficient tool or a hammer to kill a fly? *Personnel Psychology, 55*(4), 827-849.

De Meijer, L. A. L., Born, M. Ph., Terlouw, G., & Van der Molen, H. T. (2006). Applicant and method factors related to ethnic score differences in personnel selection: A study at the Dutch police. *Human Performance, 19*(3), 219-251.

Demerouti, E., Bakker, A. B., & Schaufeli, W. B. (2005). Spillover and crossover of exhaustion and life satisfaction among dual-earner parents. *Journal of Vocational Behavior, 67*(2), 266-289.

Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71*(1), 9-15.

Ensher, E. A., Grant-Vallone, E. J., & Marelich, W. D. (2002). Effects of perceived attitudinal and demographic similarity on protégés' support and satisfaction gained from their mentoring relationships. *Journal of Applied Social Psychology, 32*(7), 1407-1430.

Ensher, E. A., & Murphy, S. E. (1997). Effects of race, gender, perceived similarity, and contact on mentor relationships. *Journal of Vocational Behavior, 50*(3), 460-481.

Ettenson, R., Shanteau, J., & Krogstad, J. (1987). Expert judgment: Is more information better? *Psychological Reports, 60*, 227-238.

Ferris, G. R., & Judge, T. A. (1991). Personnel/human resources management: A political influence perspective. *Journal of Management, 17*(2), 447-488.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York: McGraw-Hill.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*(4), 652-670.

Gamliel, E., & Cahan, S. (2007). Mind the gap: Between-group differences and fair test use. *International Journal of Selection and Assessment, 15*(3), 273-282.

Goldstein, H. W., Yusko, K. P., Braverman, E. P, Smith, D. B., & Chung, B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology, 51*(2), 357-374.

Goldstein, H. W., Yusko, K. P., & Nicolopoulos, V. (2001). Exploring black-white subgroup differences of managerial competencies. *Personnel Psychology, 54*(4), 783-807.

Goldstein, H. W., Zedeck, S., & Goldstein, I. L. (2002). *G*: Is this your final answer? *Human Performance, 15*(1/2), 123-142.

Gorman, C. D., Clover, W. H., & Doherty, M. E. (1978). Can we learn anything about interviewing real people from "interviews" of paper people? Two studies and the external validity of a paradigm. *Organizational Behavior and Human Performance, 22*(2), 165-192.

Graves, L. M., & Karren, R. J. (1992). Interviewer decision processes and effectiveness: An experimental policy-capturing investigation. *Personnel Psychology, 45*(2), 313-341.

Graves, L. M., & Powell, G. N. (1995). The effect of sex similarity on recruiters' evaluations of actual applicants: A test of the similarity-attraction paradigm. *Personnel Psychology, 48*(1), 85-98.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, Ch. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30.

Hagendoorn, L. (1995). Intergroup biases in multiple group systems: The perception of ethnic hierarchies. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (pp. 199-211). London: John Wiley & Son.

Hammond, K. R., Frederick, E., Robillard, N., & Victor, D. (1989). Application of cognitive theory to the student-teach dialogue. In D. A. Evans & V. L. Patel (Eds.), *Cognitive science in medicine: Biomedical modelling* (pp. 173-210). Cambridge, MA: The MIT Press.

Hattrup, K., Rock, J., & Scalia, Ch. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology, 82*(5), 656-664.

Heneman, H. G., & Heneman, R. L. (1994). *Staffing organizations*. Middleton, WI: Mendota House.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve*. New York: Free Press Paperbacks.

Hirsh, H. Rothstein, Northrop, L. C., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology, 39*(2), 399-420.

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

Hoffmann, P. J. (1960). The paramorphic representation in clinical judgment. *Psychological Bulletin, 47*(2), 116-131.

Hofman, W. H. A. (1993). *Effectief onderwijs aan allochtone leerlingen* [Teaching ethnic minority pupils effectively]. Doctoral dissertation. Delft, The Netherlands: Eburon.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: *G*c-*G*f theory. In D. P. Flanagan & J. L. Genshaft (Ed.), *Contemporary intellectual assessment: theories, tests, and issues*. New York: Guilford Press.

Hough, L. M. (1998). Personality at work: Issues and evidence. In M. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 131-159). Hillsdale, NJ: Erlbaum.

Hoving, I., Dibbits, H., & Schrover, M. (2005).). *Cultuur en migratie in Nederland. Veranderingen van het alledaagse 1950-2000* [Culture and migration in The Netherlands. Changes of everyday life 1950-2000]. The Hague, The Netherlands: Sdu Uitgevers.

Hraba, J., Hagendoorn, L., & Hagendoorn, R. (1989). The ethnic hierarchy in The Netherlands: Social distance and social representation. *British Journal of Social Psychology, 28*(1), 57-69.

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83*(2), 179-189.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitude, job knowledge, and job performance. *Journal of Vocational Behavior, 29*(3), 340-362.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86*(4), 721-736.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67*(5), 577-580.

Jencks, C. (1998). Racial bias in testing. In C. Jencks, & M. Phillips (Eds.), *The Black-White test score gap* (pp. 55-85). Washington, DC: Brookings Inst.

Jensen, A. R. (1993). Why is reaction time correlated with psychometric *g*? *Current Directions in Psychological Science, 2*(2), 53-56.

Jones, J. W., Brasher, E. E., & Huff, J. W. (2002). Innovations in integrity-based personnel selection: Building a technology-friendly assessment. *International Journal of Selection and Assessment, 10*(1/2), 87-97.

Jussim, L., (1990). Social reality and social problems: The role of expectancies. *Journal of Social Issues, 46*(2), 9-34.

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review, 98*(1), 54-73.

Jussim, L. (1993). Accuracy in interpersonal expectations: A reflection-construction analysis of current and classic research. *Journal of Personality, 61*(4), 637-668.

Jussim, L., Coleman, L. M., & Lerch, L. (1987). The nature of stereotypes: A comparison and integration of three theories. *Journal of Personality and Social Psychology, 52*(3), 536-546.

Jussim, L., Fleming, C.J., Coleman, L., & Kohberger, C. (1996). The nature of stereotypes II: A multiple-process model of evaluations. *Journal of Applied Social Psychology, 26*(4), 283-312.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*(4), 657-690.

Karren, R. J., & Barringer, M. W. (2002). A review and analysis of the policy-capturing methodology in organizational research: guidelines for research and practice. *Organizational Research Methods, 5*(4), 337-361.

Kinicki, A. J., Lockwood, C. A., Hom, P. W., & Griffeth, R. W. (1990). Interviewer predictions of applicant qualifications and interviewer validity: Aggregated and individual analyses. *Journal of Applied Psychology, 75*(5), 477-486.

Klimoski, R. J., & Donahue, L. M. (2001). Person perception in organizations: An overview of the field. In M. London (Ed.), *How people evaluate others in organizations* (pp. 5-43). Mahwah, NJ: Lawrence Erlbaum.

Klinkenberg, E. L., & Van Leeuwen, A. E. (2003). *Voortgangsverslag ontwikkeling M5Q-IWSP* [Progress report development M5Q-IWSP]. Culemborg, The Netherlands: Meurs Personeelsadvies.

Kunda, Z. (1999). *Social cognition. Making sense of people.* Cambridge, MA: The MIT Press.

Landelijk Expertise Centrum Diversiteit (LECD; 2006). *Jaarverslag 2005* [Annual report 2005]. Apeldoorn, The Netherlands: LECD.

Lane, D. M., Murphy, K. R., & Marques, T. E. (1982). Measuring the importance of cues in policy capturing. *Organizational Behavior and Human Performance, 30*(2), 231-240.

Lankau, M. J., Riordan, Ch. M., & Thomas, Ch. H. (2005). The effects of similarity and liking in formal relationships between mentors and protégés. *Journal of Vocational Behavior, 67*(2), 252-265.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research, 39*(2), 329-358.

Lee, K., Ashton, M. C., & De Vries, R. E. (2005). Predicting workplace delinquency and integrity with the HEXACO and five-factor models of personality structure. *Human Performance, 18*(2), 179-197.

Lem, J., & Van Doorn E. (2000). *Voortgangsrapportage. Onderzoek "kenmerkende voorspellers" politie* [Progress report. Study "noticeable predictors" police]. Culemborg, The Netherlands: Meurs Personeelsadvies.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admission: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90*(3), 442-452.

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91*(5), 1181-1188.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.* Glenview, IL: Scott, Foresman, and Company.

Linville, P. W. (1982). The complexity-extremity effect and age-based stereotyping. *Journal of Personality and Social Psychology, 42*(2), 193-211.

Linville, P. W., & Jones, E. E. (1980). Polarized appraisals of out-group members. *Journal of Personality and Social Psychology, 38*(5), 689-703.

Locksley, A., Borgida, E., Brekke, N., & Hepburn, C. (1980). Sex stereotypes and social judgment. *Journal of Personality and Social Psychology, 39*(5), 821-831.

Locksley, A., Hepburn, C., & Ortiz, V. (1982a). On the effects of social stereotypes on judgments of individuals: A comment on Grant and Holmes's "The integration of implicit personality theory schemas and stereotypic images". *Social Psychology Quarterly, 45*(4), 270-273.

Locksley, A., Hepburn, C., & Ortiz, V. (1982b). Social stereotypes and judgments of individuals: An instance of base-rate fallacy. *Journal of Experimental Social Psychology, 18*(1), 23-42.

Lucassen, J. (2005). Appendix. Een kort overzicht van de immigratie naar Nederland in de twintigste eeuw [Appendix. A short overview of the immigration into The Netherlands in the 20th century]. In I. Hoving, H. Dibbits, & M. Schrover (Eds.), *Cultuur en migratie in Nederland. Veranderingen van het alledaagse 1950-2000* [Culture and migration in The Netherlands. Changes of everyday life 1950-2000] (pp. 415-428). The Hague, The Netherlands: Sdu Uitgevers.

Lucassen, J., & Penninx, R. (1994). *Nieuwkomers, nakomelingen, Nederlanders. Immigranten in Nederland 1550-1993* [Newcomers, descendants, Dutchmen. Immigrants in The Netherlands 1550-1993]. Amsterdam: Het Spinhuis.

Markus, H., Smith, J., & Moreland, R. L. (1985). Role of the self-concept in the perception of others. *Journal of Personality and Social Psychology, 49*(6), 1494- 1512.

McCroskey, J. C., Richmond, V. P., & Daly, J. A. (1975). The development of a measure of perceived homophily in interpersonal communication. *Human Communication Research, 1*, 323-332.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63-91.

McDaniel, M. A., Morgeson, F. P., Bruhn Finnegan, E., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*(4), 730-740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*(1/2), 103-113.

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33*(5), 515-525.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*(4), 599-616.

McFarland, L. A., Ryan, A. M., Sacco, J. M., & Kriska, S. D. (2004). Examination of structured interview ratings across time: The effects of applicant race, rater race, and panel composition. *Journal of Management, 30*(4), 435-452.

McIntyre, M. D., & James, L. R. (1995). The inconsistency with which raters weight and combine information across targets. *Human Performance, 8*(2), 95-111.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*(3), 683-729.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*(4), 1029-1049.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R. et al. (1992). Studies of the structural behavioral interview. *Journal of Applied Psychology, 77*(5), 571-587.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*(6), 640-647.

Murphy, K. R. (2002). Can conflicting perspectives on the role of *g* in personnel selection be resolved? *Human Performance, 15*(1/2), 173-186.

Murphy, K. R., & Dzieweczynski, J. L. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance? *Human Performance, 18*(4), 343-357.

Naeyé, J., Huberts, L., Van Zweden, C., Busato, V., & Berger, B. (2004). *Integriteit in het dagelijkse politiewerk* [Integrity during daily police work]. Amsterdam: Vrije Universiteit.

Nguyen, N. T., & McDaniel, M. A. (2003). Response instructions and racial differences in a situational judgment test. *Applied H.R.M. Research, 8*(1), 33-44.

Nguyen, N. T., McDaniel, M. A., & Whetzel D. L. (2005, April). *Subgroup differences in situational judgment test performance: A meta-analysis.* Paper presented at the 20th Annual Conference of the Society of Industrial and Organizational Psychology, Los Angeles, CA.

Nijsten, C. (1998). *Opvoeden in Turkse gezinnen in Nederland* [Raising in Turkish families in The Netherlands]. Assen, The Netherlands: Van Gorcum.

*Occupational Information Network (O*NET) OnLine developed for the US Department of Labor by the National O*NET Consortium.* (2007, January 31). Retrieved January 31, 2007, from http://online.onetcenter.org

*Occupational Information Network (O*NET) OnLine developed for the US Department of Labor by the National O*NET Consortium.* (2007, May 22). Retrieved May 22, 2007, from http://online.onetcenter.org

O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L., III., & Lawrence, A. (2007). Incremental validity of situational judgment tests

for task and contextual performance. *International Journal of Selection and Assessment, 15*(1), 19-29.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79*(6), 845-851.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). An interactive video assessment of conflict resolution skills. *Personnel Psychology, 51*(1), 1-24.

Ones, D. S. (1993). *The construct validity evidence for integrity tests.* Unpublished doctoral dissertation, University of Iowa, Iowa City.

Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology, 75*(3), 255-276.

Ones, D. S., & Viswesvaran, C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale job applicant data sets. *Journal of Applied Psychology, 83*(1), 35-42.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance [Monograph]. *Journal of Applied Psychology, 78*(4), 679-703.

Outtz, J. L. (2002). The role of cognitive ability tests in employment selection. *Human Performance 15*(1/2), 161-171.

Pinto, D. (2004). *Beeldvorming en integratie. Is integratie het antwoord?* [Representation and integration. Is integration the answer?]. Houten, The Netherlands: Bohn Stafleu Van Loghum.

Plomin, R. (1988). The nature and nurture of cognitive abilities. In R.J. Sternberg (ed.), *Advances in the psychology of human intelligence* (Vol. IV). Hilllsdale, NJ: Erlbaum.

Politieacademie (2007). Verschil moet er zijn [Things differ]. *Politieacademiekrant, 9*, 2007.

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

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology, 74*(5), 770-780.

Pynes, J. E., & Bernardin, H. J. (1989). Predictive validity of an entry-level police officer assessment center. *Journal of Applied Psychology, 74*(5), 831-833.

Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior knowledge in complex training performance. *Journal of Applied Psychology, 80*(6), 721-730.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44*(2), 321-332.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*(6), 880-887.

Rijks Psychologische Dienst (1975). *Politie Intelligentie Test* [Police Intelligence Test]. The Hague, The Netherlands: Rijks Psychologische Dienst (RPD).

Robertson, I. T., & Smith, M. (2001). Personnel selection. *Journal of Occupational and Organizational Psychology, 74*(4), 441-472.

Rokeach, M., & Mezel, L. (1966). Race and shared belief as factors in social choice. *Science, 151*(3707), 167-172.

Roth, P. L., Huffcutt, A. I., & Bobko, Ph. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88*(4), 694-706.

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., & Bobko, P. (2002). Correction for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology, 87*(2), 369-376.

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

Ryan, A. M. (2001). Explaining the black-white test score gap: The role of test perceptions. *Human Performance, 14*(1), 45-75.

Sacco, J. M., Scheu, C. R., Ryan, A. M., & Schmitt, N. (2003). An investigation of race and sex similarity effects in interviews: A multilevel approach to relational demography. *Journal of Applied Psychology, 88*(5), 852-865.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52*, 471-499.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., De Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology, 88*(6), 1068-1081.

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability and job knowledge. *Current Directions in Psychological Science, 2*(1), 8-9.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*(1), 162-173.

Schmitt, N., & Mills, A. E. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 86*(3), 451-458.

Segall, M. H., Dasen, P. R., Berry, J. W., & Poortinga, Y. H. (1999). *Human behavior in global perspective. An introduction to cross-cultural psychology* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Shanteau, J. (1991). Psychological characteristics and strategies of experts. In G. Wright, & F. Bolger (Eds.), *Expertise and decision support*. New York: Plenum.

Shanteau, J., Grier, M., Johnson, J., & Berner, E. (1991). Teaching decision-making skills to student nurses. In J. Baron & R. V. Brown (Eds.), *Teaching decision making to adolescents*. Hillsdale, NJ: Erlbaum.

Singer, M. S., & Bruhns, C. (1991). Relative effect of applicant work experience and academic qualification on selection interview decisions: A study between sample generalizability. *Journal of Applied Psychology, 76*(4), 550-559.

Stewart, T. R. (1988). Judgment analysis: Procedures. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment. The SJT view* (pp. 41-74). Amsterdam: North Holland.

Strauss, J. P., Barrick, M. R., & Connerley, M. L. (2001). An investigation of personality similarity effects (relational and perceived) on peer and supervisor ratings and the role of familiarity and liking. *Journal of Occupational and Organizaitonal Psychology, 74*(5), 637-657.

Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin, 105*(3), 409-429.

Tajfel, H. (1978). *Differentiation between social groups: Studies in the social psychology of intergroup relations*. New York: Academic Press.

Tajfel, H. (1982). *Social identity and intergroup relations*. London: Cambridge University Press.

Te Nijenhuis, J. (1997). *Comparability of test scores for immigrants and majority group members in the Netherlands*. Unpublished doctoral dissertation,Vrije Universiteit, Amsterdam.

Te Nijenhuis, J., & Van der Flier, H. (2000). Differential prediction of immigrant versus majority group training performance using cognitive ability and personality measures. *International Journal of Selection and Assessment, 8*(2), 54-60.

Te Nijenhuis, J., Van der Flier, H., & Van Leeuwen, L. (1997). Comparability of personality test scores for immigrants and majority group members: Some Dutch findings. *Personality and Individual Differences, 23*(5), 849-859.

Turban, D. B., Dougherty, T. W., & Lee, F. K. (2002). Gender, race, and perceived similarity effects in developmental relationships: The moderating role of relationship duration. *Journal of Vocational Behavior, 61*(2), 240-262.

Turner, J. C. (1987). *Rediscovering the social group: A self-categorization theory*. Oxford, Great Britain: Blackwell.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Van den Berg, R. H. (2001). *Psychologisch onderzoek in een multiculturele samenleving* [Psychological research in a multicultural society]. Unpublished doctoral dissertation, Vrije Universiteit, Amsterdam.

Van den Berg, R. H., & Van Leest, P. (1999). Praktisch testgebruik bij allochtonen: wanneer zinnig, nuttig of waardeloos? [Practical use of tests with ethnic minorities: when useful, useless, or worthless?]. *Psycholoog, juni*, 256-260.

Van der Maesen, P. E. A. M. (1992). *Het rendement van personeelsselectie* [The efficiency of personnel selection]. Unpublished doctoral dissertation, Rijksuniversiteit Groningen, The Netherlands.

Van der Velden, R. K. W. (1991). *Sociale herkomst en schoolsucces* [Socio-economic status and school success]. Unpublished doctoral dissertation, Rijksuniversiteit Groningen, The Netherlands.

Van Emmerik, I. H., & Jawahar, I. M. (2006). The independent relationships of objective and subjective workload with couples' mood. *Human Relations, 59*(10), 1371-1392.

Van Herk, H., Poortinga, Y. H., & Verhallen, Th. M. M. (2004). Response styles in rating scales. Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*(3), 346-360.

Van Iddekinge, C. H., Taylor, M. A., & Eidson, C. E., Jr. (2005). Broad versus narrow facets of integrity: predictive validity and subgroup differences. *Human Performance, 18*(2), 151-177.

Van Leest, P. (1997). *Persoonlijkheidsmeting bij allochtonen* [Ethnic minority personality measurement]. Unpublished doctoral dissertation, Vrije Universiteit, Amsterdam.

Van Leeuwen, A. E. (2000). *Constructie van de M5Q voor IWSP* [Construction of the M5Q for IWSP]. Culemborg, The Netherlands: Meurs Personeelsadvies.

Van Loon, H. (2003). *De praktijkopdracht van het IWSP* [The AC of the IWSP] (Internal publication). Apeldoorn, The Netherlands.

Van Rijn, A. S., Zorlu, A., Bijl, R. V., & Bakker, B. F. M. (2004). *De ontwikkeling van een integratiekaart* [The development of an integration map]. The Hague, The Netherlands: WODC & CBS.

Verkuyten, M., Hagendoorn, L., & Masson, K. (1996). The ethnic hierarchy among majority and minority youth in The Netherlands. *Journal of Applied Social Psychology, 26*(12), 1104-1118.

Vogel, J. (2005). *Cultuur en migratie in Nederland. Nabije vreemden* [Culture and migration in The Netherlands. Near strangers]. The Hague, The Netherlands: Sdu Uitgevers.

Wahlstrom, R., Hummers-Pradier, E., Lundborg, C. S., Muskova, M., Lagerlov, P., et al. (2002). Variations in asthma treatment in five European countries – Judgment analysis of case simulations. *Family Practice, 19*(5), 452-460.

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

Wallace, H. A. (1923). What is in the corn judge's mind? *Journal of the American Society of Agronomy, 15*, 300-304.

Wanek, J. E. (1999). Integrity and honesty testing: What do we know? How do we use it? *International Journal of Selection and Assessment, 7*(4), 183-195.

Weekley, J. A., & Jones, C. (1997). Video-based situational judgment testing. *Personnel Psychology, 50*(1), 25-49.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*(3), 679-700.

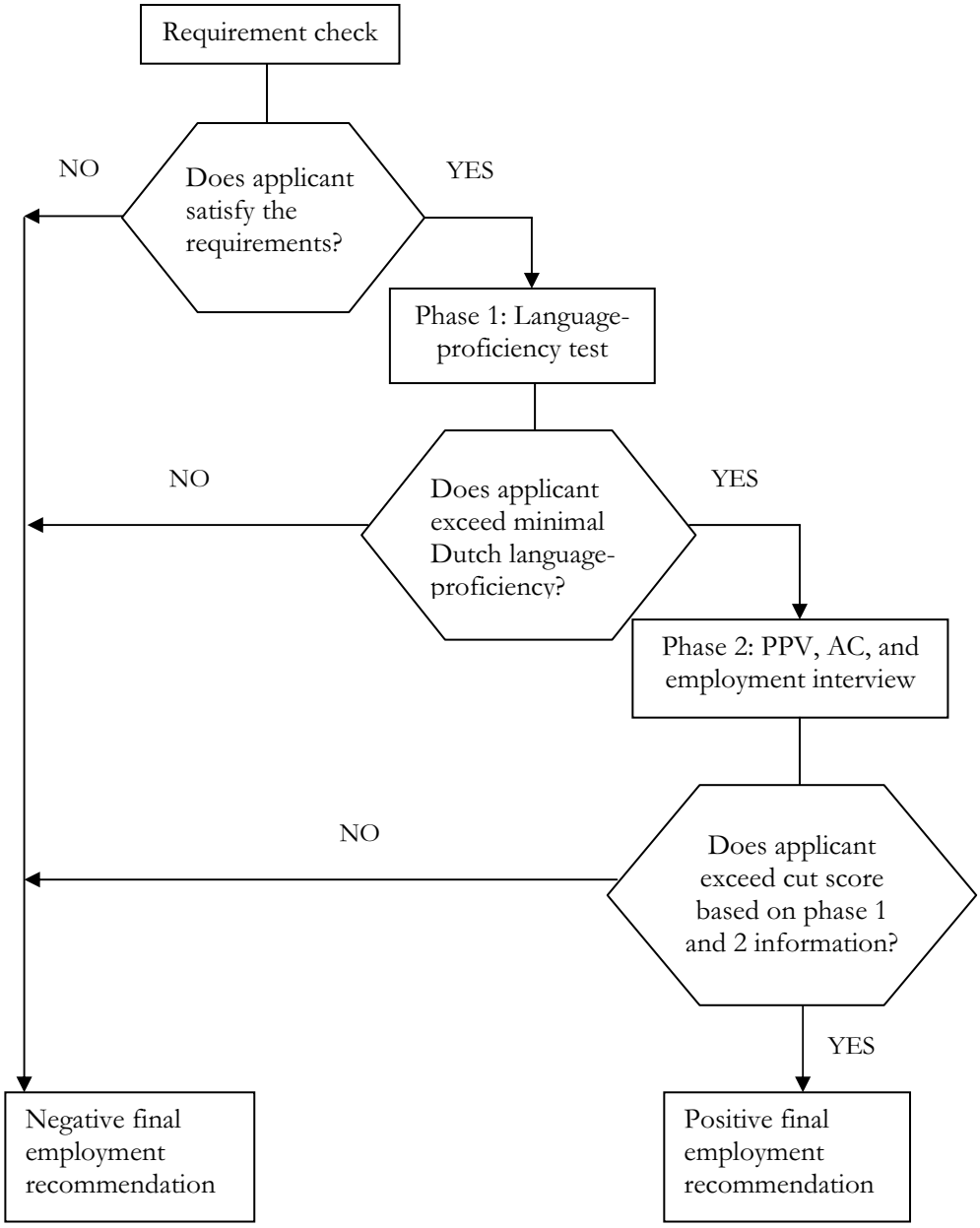
Weijters, G., & Scheepers, P. (2003). Verschillen in sociale integratie tussen etnische groepen: beschrijving en verklaring [Differences between ethnic groups in social integration: Description and explanation]. *Mensch en Maatschappij, 78*(2), 144-157.

Wiggins, J. S., & Trapnell, P. D. (1996). A dyadic-interactional perspective on the five-factor model. In J.S. Wiggins (Ed.), *The five-factor model of personality* (pp. 88-162). New York: The Guilford Pres.

Zedeck, S., Tziner, A., & Middlestadt, S. E. (1983). Interview validity and reliability: An individual analysis approach. *Personnel Psychology, 36*(2), 355-370.

Appendices

Appendix A



Note. PPV = personality questionnaire

Figure. Selection process at the Center of Competence Assessment and Monitoring (CCM) of the Police Academy of The Netherlands

Appendix B

Dimensions, Dimension Descriptions, and Selection Tool Used

Dimension	Description	Instrument
Communication Skills	The ability to transmit information, ideas, and opinions, both verbally and non-verbally.	AC, Interview
Social Skills	The desire to have and begin social contacts, and to keep up these contacts.	AC, Interview
Empathy	The ability to put oneself in the thoughts, feelings, and reactions of others.	AC
Initiative	Taking or starting action of one's own accord, without incitement from outside, instead of waiting.	AC
Flexibility	Changing tasks fast and easily, being able to adapt to changing circumstances, and desiring changes and variation.	Interview
Emotional Stability	Being able to cope with emotional far-reaching situations.	Interview
Stress Tolerance	Being able to cope with high work- and time-pressure in daily work situations.	AC, Interview
Authority	Being able to influence others, both verbally and non-verbally, and being accepted as an authority by other people.	AC
Decisiveness	Being able and prepared to make decisions in dilemmas and with incomplete information, and taking responsibility for the consequences of these decisions.	AC
Tolerance Towards Others	Accepting and respecting differences between people, and taking these differences into consideration in one's own behavior.	Interview
Integrity	Being aware of the general acknowledged norms and values in society and showing willingness to act on these.	Interview
Self-Understanding	Being aware of one's own qualities and behavior, being able to reflect on qualities and behavior, and willing to integrate these reflections in future behavior.	Interview

Appendix C

Example of Integrity-SJT Item

Description of situation:
A police officer (police officer 1) comes to work on his motorbike. When he enters the parking garage of the police station he accidentally hits a police car, causing a big scratch on the police car. Shortly after, he meets a colleague (police officer 2) and tells her what happened.

Police officer 1:
"Hi! Listen: A just entered the parking garage with my motorbike and caused a big scratch on one of the police cars. I feel really bad about it and, actually, I don't know what to do."

Possible reactions of police officer 2:
1. Don't worry about it! Police cars are covered with scratches. (Factor 1)
2. O... I'm sorry. If I were you, I would report it to the chief. (Factor 3)
3. Well, that's pretty stupid of you!! You have to report it to the chief! (Factor 2)
4. The only thing you can do is to report it to the chief! And if you're not going to do it, I will!! (Factor 4)

Appendix D

Integrity-Related Dimensions and their Descriptions

Dimension	Description
Employment Interview and Final Recommendation	
Integrity	Being aware of the general acknowledged norms and values in society and showing willingness to act on these.
In-Depth Interview	
Modesty	Being modest, unassuming, and seeing oneself as an ordinary person without any claim to special treatment.
Honesty	Being genuine in interpersonal relations and unwilling to manipulate others.
Morality	Being able to avoid fraud and corruption and unwilling to take advantage of other individuals or of society at large.
Avoidance of Materialism	Being uninterested in possessing lavish wealth, luxury goods, and signs of high social status.
How-I-Think Questionnaire (HIT)	
Self-Centered	According status to one's own view, expectations, needs, rights, immediate feelings, and desires to such a degree that the legitimate views, etc., of others are scarcely considered or are disregarded altogether.
Blaming Others	Misattributing blame to outside sources or misattributing blame for one's victimization or other misfortune to innocent others.
Minimizing/ Mislabeling	Depicting antisocial behavior as causing no real harm or referring to others with a belittling or dehumanizing label.
Opposition-Defiance	Being disrespectful for rules, laws, or authorities.

Note. Definitions of the facets of the in-depth Integrity interview are from Lee & Ashton (2004) and definitions of the (sub-) dimensions of the How-I-Think questionnaire (HIT) are from Barriga et al. (2001). Definitions of the sub-dimensions Physical Aggression, Stealing, and Lying of the HIT were not listed here, because we assumed that they are self-explanatory.

Dankwoord

Een proefschrift schrijf je niet alleen. Aan deze dissertatie hebben bijna dertigduizend mensen meegewerkt, aan wie ik veel dank verschuldigd ben.

Speciaal wil ik bedanken, mijn promotoren Marise Born en Henk van der Molen. Ik had me geen betere mentoren kunnen wensen. Marise, hartelijk dank voor al jouw hulp! Jij hebt ervoor gezorgd dat er een flinke vaart in het proces bleef en jij hebt me steeds weer gestimuleerd om kritisch na te denken over mijn werk en mijn plannen. Henk, jij was van bijzonder belang voor de *helicopter view* als ik weer eens tot over mijn oren in de details zat. En ondanks jullie overvolle agenda's, hadden jullie altijd tijd voor overleg, begeleiding en motivatie. Dank ook aan Henk Schmidt, Fons van de Vijver en Jasper von Grumbkow voor hun bereidheid zitting te nemen in de leescommissie.

Dank aan mijn externe begeleidingsteam bij de Nederlandse Politieacademie in Apeldoorn, Jaap van Zielst, Gert Terlouw en Hans van Loon. Zonder de hulp van het Centrum voor Competentiemeting en Monitoring (CCM) van de Nederlandse Politieacademie en zonder jullie toewijding had dit proefschrift niet tot stand kunnen komen. Ik wil jullie hartelijk danken voor jullie betrokken houding, het vertrouwen dat jullie mij gegeven hebben om het onderzoek uit te voeren en jullie wijze adviezen en deskundigheid. In een politiewereld die ik niet kende, hebben jullie me fantastisch wegwijs gemaakt!

Dank aan de deskundigen met wie ik het genoegen heb gehad samen te mogen werken. In Den Haag: Paul van der Maesen, dank voor de hulp en adviezen bij de ontwikkeling van de *Situational Judgment Test* (SJT). In Arnhem: Sonia Sjollema van de Nederlandse Stichting voor Psychotechniek (NSvP): dank voor de financiële hulp die geboden werd in het kader van de ontwikkeling van de SJT. In Apeldoorn en Amsterdam: de selectiepsychologen, –adviseurs en testleiding van het CCM, dank voor jullie inzet bij de dataverzameling van verschillende studies en Hellen Westerveld, dank voor het harde werk in de beginfase van de ontwikkeling van de SJT. In Rotterdam: Miriam Heemskerk: dank voor het consciëntieuze werk in een later stadium van de ontwikkeling van de SJT. In Amsterdam: Peter Dekker, dank voor de statistische hulp. In het buitenland: Iain Coyne, thank you for helping me with my academic writing in the beginning of my Ph.D.-project.

Dank aan mijn (ex-)collega's bij het Instituut voor Psychologie aan de Erasmus Universiteit Rotterdam voor het creëren van de fijne werksfeer: Lyanda, Nevra, Stefan, Annet, Lydia, Eva, Heleen, Edwin, Arnold, Marjan, Alec, Daantje, René, Despoina, Benjamin, Gera, Janneke, Maria, en Annemarie en op het secretariaat: Hanny en Mirella.

Dank aan álle kandidaten die deel hebben genomen aan de selectieprocedure van de Nederlandse Politieacademie, zonder wie het onderzoek niet had kunnen plaatsvinden. De politie-experts die in het kader van de ontwikkeling van de SJT benaderd zijn: bedankt voor de bereidheid om mee te doen.

Dank aan mijn paranimfen, die mij, ieder op een eigen wijze hebben bijgestaan in de tocht naar de eindstreep. Martine, dank voor jouw luisterend oor en relativerende humor. Je weet mij altijd weer te begrijpen en goed aan te voelen. Sofie, in de afgelopen jaren ben jij uitgegroeid tot veel meer dan een collega. Je bent mijn sportmaatje, reisgenoot en vriendin. Ik voel me vereerd dat jullie me ook in dit laatste stadium van het promotietraject terzijde willen staan!

Dank aan de mensen van het thuisfront: Willem en Laura, Guus en Marijke, Annemieke, Wiebren, Nienke en Marit, Karlijn en Peter, Marieke, Noor, Sacha en Eric, dank voor alle steun en belangstelling. Mijn ouders, Wim en Dimphe, voor alle stimulans die jullie mij gegeven hebben. Jullie hebben me geleerd om hard te werken en het beste uit mezelf te halen. Daar ben ik jullie zeer dankbaar voor. Sjoerd, zonder jouw relativeringsvermogen en steun bij alles was dit proefschrift er nooit gekomen! En natuurlijk dank aan mijn kleine Stijn, omdat je mijn lieve zoon bent.

Dank aan iedereen.

Lonneke
Rotterdam, 2008

Kurt Lewin Institute Dissertation Series

The "Kurt Lewin Institute Dissertation Series" started in 1997. Since 2006 the following dissertations have been published:

2006-1: Maria Dijkstra: *Workplace Conflict and Individual Well-Being*
2006-2: Ruud Custers: *On the underlying mechanisms of nonconscious goal pursuit*
2006-3: Ellen Dreezens: *The missing link: the relationship between values and attitudes*
2006-4: Jacquelien van Stekelenburg: *Promoting or preventing social change. Instrumentality, identity, ideology and groupbased anger as motives of protest participation*
2006-5: Huadong Yang: *Siding in a conflict in China and in the Netherlands*
2006-6: Tomas Ståhl: *Determinants of Fairness-based and Favorability-based Reactions to Authorities' Decisions*
2006-7: Astrid Homan: *Harvesting the value in diversity: Examining the effects of diversity beliefs, cross-categorization, and superordinate identities on the functioning of diverse work groups*
2006-8: Saskia Schwinghammer: *The Self in Social Comparison*
2006-9: Carmen Carmona Rodríguez: *Inferior or Superior: Social Comparison in Dutch and Spanish Organizations*
2006-10: Martijn van Zomeren: *Social-psychological paths to protest: An integrative perspective*
2007-1: Nils Jostmann: *When the going gets tough… How action versus state orientation moderates the impact of situational demands on cognition, affect, and behavior*
2007-2: Belle Derks: *Social identity threat and performance motivation: The interplay between ingroup and outgroup domains*
2007-3: Helma van den Berg: *Feeling and Thinking in Attitudes*
2007-4: Karin C.A. Bongers: *You can't always get what you want! Consequences of success and failure to attain unconscious goals*
2007-5: Lotte Scholten: *Motivation matters: Motivated information processing in group and individual decision-making*
2007-6: Debra Trampe: *Social influence: Social comparison, construal, and persuasion processes*
2007-7: Clemens Wenneker: *Processes underlying biased language use*
2007-8: Yaël de Liver: *Ambivalence: on the how and when of attitudinal conflict*
2007-9: Erik de Kwaadsteniet: *Uncertainty in social dilemmas*
2007-10: Hugo Alberts: *Processes of self-control and ego depletion*

2007-11: Loran Nordgren: *Thinking about Feeling: The Nature and Significance of the Hot/Cold Empathy Gap*
2007-12: Stefan Thomas Mol: *Crossing Borders with Personnel Selection from expatriates to multicultural teams*
2007-13: Hilbrand Oldenhuis: *I know what they think about us: Metaperceptions and intergroup relations*
2007-14: Arnaud Wisman: *New Directions in Terror Management Theory*
2007-15: Gert Homsma: *Making Errors Worthwhile: Determinants of Constructive Error Handling*
2007-16: Elianne van Steenbergen: *Work-Family Facilitation: A Positive Psychological Perspective on Role Combination*
2007-17: Unna Danner: *By Force of Habit: On the Formation and Maintenance of Goal-Directed Habits*
2007-18: Maureen Tumewu: *The Social Psychology of Gender Differences and Procedural Justice in Close Relationships*
2008-1: Marijke van Putten: *Dealing with missed opportunities. The causes and boundary conditions of inaction inertia*
2008-2: Marjolein Maas: *Experiential Social Justice Judgment Processes*
2008-3: Lonneke de Meijer: *Ethnicity effects in police officer selection: Applicant, assessor, and selection-method factors*