# Investigation of Examiner Effects on Test Takers in Standardized Achievement Tests with Special Regard to Gender

Dissertation zur Erlangung des akademischen Grades
Doktorin der Philosophie (Dr. phil.)

am Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

vorgelegt von Mag. Isabella Vormittag

Berlin, 2011

Erstgutachterin: Prof. Dr. Tuulia M. Ortner

Zweitgutachterin: Prof. Dr. Bettina Hannover

Tag der Disputation: 27.10.2011

# Table of Contents

# Danksagung

For reasons of data protection, the acknowledgements are not included in this version.

# Abstract

In educational and vocational settings standardized tests and questionnaires are widely used (Fernández-Ballesteros, 1999). Procedures vary between specific assessment methods considerably by covering paper-pencil and computerized assessment as well as testing of groups or individuals. With focus on standardized tests, in most situations two interacting parts can be differentiated: a test taker – who is requested to show a certain behavior – and an examiner – who administers the procedure (e.g., Anastasi & Urbina, 1997). Research has repeatedly shown that the examiner can have a significant influence on the behavior of the test taker (cf. Rosenthal, 1976; Rosenthal, 1995; Sattler & Theye, 1967). Whereas many of those undesired effects can be avoided through standardization of assessment (e.g., Sattler & Theye, 1967), some sources of examiner effects seem to be inevitable. At special risk are physical cues that are related to social roles and stereotypes, e.g. gender and ethnicity. Research has shown that examiners' ethnicity may influence the performance of test takers on standardized achievement tests (Huang, 2009; Mishra, 1980). There is a lack of current research considering possible gender effects in such situations. The major aim of this thesis was to investigate the influence of the examiner's gender on the test taker's performance on standardized tests.

In achievement testing test taker usually cannot choose the person who will administer the test. We therefore do not know if test takers would prefer certain examiners by implicitly expecting better or more convenient test conditions. There is currently no investigation of test takers' perception of and preference for examiners. Therefore, a second aim of this thesis was to explore how test takers perceive and rate examiners, and moreover who they prefer for administration if they are given a choice.

This doctoral thesis comprises four studies. The first and the second studies applied an individual face-to-face testing at the Free University of Berlin. A verbal knowledge test consisting of two modules was employed: The first part measures self-estimated verbal knowledge and the second part measures amount of de facto verbal knowledge. Test takers were nonpsychology students ($N$ = 93 in Study 1; $N$ = 114 in Study 2), participating voluntarily. Examiners were psychology students – either diploma students ($N$ = 20, Study 1) or diploma and bachelor students ($N$ = 22, Study 2). The results of the first study revealed that male and female test takers estimated their own knowledge higher when tested by a female examiner. In the second study additionally

perceived attractiveness of the examiner was included and a significant three-way interaction of gender of test taker, gender of examiner, and perceived attractiveness of the examiner on the performance found: Test takers who were tested by an attractive same-gender examiner showed poorer performance than test takers in mixed-gender settings or test takers who perceived the examiner as not attractive.

The third study investigated how examiners are perceived. First, a pilot study was conducted, asking test takers ($N$ = 129) to choose either a male or a female examiner for an upcoming testing. Significantly more test takers decided for a female than for a male examiner. In the following main study an online design was employed. Students ($N$ = 375) from different universities in Germany watched four short video clips of male and female examiners of two age groups giving standardized test instructions. Participants were asked to rate the examiners' expertise and social competence and eventually choose one favorite examiner. Results showed no differences in perceived expertise due to gender, but higher ratings of social competence for female examiners. Women were significantly more often preferred than men.

The fourth study used data from the German Socio-economic panel. The sample consisted of 2,863 participants who took part in an additional short achievement test measuring perceptual speed. The test was applied via laptop with one of 178 examiners present. Multilevel analyses revealed that test taker's age and examiner's gender were significant predictors of the performance. This study showed – albeit small – examiner gender effects in a large representative German sample of participants with different ethnic and education background.

Summarizing, the results of the four studies showed that the examiner influences test takers even in standardized testing. First, self-estimations and expectations towards the own achievement as well as actual performance seemed to be affected by the examiner gender. Second, results indicated that stereotypical perceptions led to different prospect of the assessment. In the thesis an integrating discussion of the four studies is presented where practical implication and claims concerning future research are described.

**References:**

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River (NJ): Prentice Hall.

Fernández-Ballesteros, R. (1999). Psychological assessment: Future challenges and progresses. *European Psychologist, 4*, 248-262. doi: 10.1027//1016-9040.4.4.248

Huang, M. H. (2009). Race of the interviewer and the black-white test score gap. *Social Science Research, 38*, 31-40. doi: 10.1016/j.ssresearch.2008.07.004

Mishra, S. P. (1980). Influence of Examiners Ethnic Attributes on Intelligence-Test Scores. *Psychology in the Schools, 17*, 117-122. doi: 10.1002/1520-6807(198001)17:1<117::AID-PITS2310170122>3.0.CO;2-6

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research*. New York, N.Y.: Irvington Publishers, Inc.

Rosenthal, R. (1995). Critiquing Pygmalion: a 25-Year Perspective. *Current Directions in Psychological Science, 4*, 171-172. doi: 10.1111/1467-8721.ep10772607

Sattler, J. M., & Theye, F. (1967). Procedural, Situational, and Interpersonal Variables in Individual Intelligence Testing. *Psychological Bulletin, 68*, 347-360. doi: 10.1037/h0025153

Socio-Economic Panel (SOEP) data for the years 1984-2009, version 26. (2010).

# Zusammenfassung

Im Berufs- wie Ausbildungskontext ist die Anwendung psychologischer Tests und Fragebögen weit verbreitet (Fernández-Ballesteros, 1999). Die konkreten Testsituationen können sehr unterschiedlich gestaltet sein, beispielsweise hinsichtlich der Verwendung von computerisierten oder Papier-Bleistift-Verfahren und bezogen auf Einzel- oder Gruppentestungen. In Hinblick auf die Vorgabe von standardisierten Tests können in den meisten Situationen zwei interagierende Seiten unterschieden werden: die Testperson, die eine bestimmte Leistung erbringen soll, und der Testleiter bzw. die Testleiterin, deren Aufgabe die Testvorgabe ist (z.B. Anastasi & Urbina, 1997). Forschungsergebnisse haben wiederholt gezeigt, dass die jeweiligen Testleiter einen bedeutsamen Einfluss auf das Verhalten der Testpersonen haben können (vgl. Rosenthal, 1976; Rosenthal, 1995; Sattler & Theye, 1967). Während einige dieser unerwünschten Effekte durch Standardisierung der Erhebung vermieden werden können (z.B. Sattler & Theye, 1967), erscheinen manche Quellen von Testleitereffekten unvermeidbar. Ein besonderes Risiko stellen visuelle Hinweisreize dar, die mit sozialen Rollen und Stereotypen assoziiert sind: beispielsweise Geschlecht oder Ethnizität. In der Forschung hat sich gezeigt, dass die ethnische Zugehörigkeit von Testleitern die Leistung von Testpersonen in standardisierten Leistungstests beeinflussen kann (Huang, 2009; Mishra, 1980). Es mangelt jedoch an aktuellen Forschungsergebnissen, die mögliche Effekte aufgrund des Geschlechts in solchen Situationen berücksichtigen. Das zentrale Ziel dieser Dissertation bestand in der Untersuchung des Einflusses vom Testleitergeschlecht auf die Leistung der Testperson in standardisierten Tests.

Bei Leistungstestungen können Testpersonen üblicherweise nicht entscheiden, wer den Test vorgeben wird. Daher wissen wir nicht, ob Testpersonen aufgrund impliziter Erwartungen an bessere oder angenehmere Testbedingungen bestimmte Testleiter bevorzugen würden. Aktuell gibt es keine Untersuchung der Wahrnehmung und Präferenz von Testpersonen in Bezug auf Testleiter. Aus diesem Grund bestand ein zweites Ziel dieser Dissertation in der Untersuchung der Wahrnehmung und Beurteilung von Testleitern durch Testpersonen. Außerdem sollten die Präferenzen der Testpersonen untersucht werden, wenn diesen die Wahl eines Testleiters bzw. einer Testleiterin ermöglicht wird.

Diese Dissertation besteht aus vier Studien. In den ersten beiden Studien wurden jeweils individuelle *face-to-face* Testungen an der Freien Universität Berlin durchgeführt. Es wurde ein verbaler Wissenstest, der aus zwei Teilen besteht, vorgegeben. Der erste Teil erfasst das selbsteingeschätzte verbale Wissen und der zweite Teil misst das tatsächliche verbale Wissen. Studierende nahmen freiwillig als Testpersonen teil ($N$ = 93 in Studie 1; $N$ = 114 in Studie 2), wobei Psychologiestudierende als Testpersonen ausgeschlossen wurden. Die Testleiter waren Psychologiestudierende, entweder aus dem Diplomstudiengang ($N$ = 20, Studie 1) oder aus Diplom- und Bachelorstudiengang ($N$ = 22, Studie 2). Die Ergebnisse der ersten Studie zeigten, dass männliche und weibliche Testpersonen ihr eigenes Wissen höher einschätzten, wenn sie von einer Testleiterin getestet wurden.

In der zweiten Studie wurde zusätzlich die wahrgenommene Attraktivität der Testleiter erhoben. Es zeigte sich, dass eine 3-fach Interaktion zwischen Testpersonengeschlecht, Testleitergeschlecht und wahrgenommener Attraktivität der Testleiter einen signifikanten Effekt auf die Leistung im Wissenstest hatte: Testpersonen, die von einem attraktiven Testleiter bzw. Testleiterin desselben Geschlechts getestet wurden, erzielten schlechtere Ergebnisse als Testpersonen in gemischtgeschlechtlichen Testsituationen oder Testpersonen, die den Testleiter bzw. die Testleiterin nicht attraktiv fanden.

In der dritten Studie wurde die Wahrnehmung von Testleitern untersucht. Zunächst wurden Testpersonen ($N$ = 129) in einer Pilotstudie gebeten, sich für eine bevorstehende Testung entweder bei einem Testleiter oder einer Testleiterin anzumelden. Es entschieden sich signifikant mehr Testpersonen für die Testleiterin als für den Testleiter. Die Hauptstudie wurde als Onlinestudie durchgeführt. Studierende ($N$ = 375) von verschiedenen Universitäten in Deutschland sahen vier kurze Videos von männlichen und weiblichen Testleitern aus zwei Altersgruppen, die eine standardisierte Testinstruktion gaben. Die Probanden waren aufgefordert die Testleiter hinsichtlich fachlicher Kompetenz und sozialer Kompetenz zu beurteilen und abschließend einen bevorzugten Testleiter bzw. Testleiterin zu wählen. Die Ergebnisse zeigten keinen Unterschied in der wahrgenommenen fachlichen Kompetenz zwischen männlichen und weiblichen Testleitern. Die soziale Kompetenz wurde bei Frauen signifikant höher eingeschätzt. Frauen wurde signifikant häufiger als bevorzugte Testleiter gewählt.

In der vierten Studie wurden Ergebnisse aus dem Sozioökonomischen Panel Deutschland verwendet. Die Stichprobe bestand aus 2,863 Probanden, die an einem

kognitiven Kurztest zur Erfassung der Wahrnehmungsgeschwindigkeit teilnahmen. Die Testung wurde am Laptop durchgeführt, wobei einer von 178 Testleitern währenddessen anwesend war. Mehrebenenanalysen zeigten, dass das Alter der Testpersonen und das Geschlecht der Testleiter einen signifikanten Einfluss auf die Leistung im Test hatten. In dieser Studie zeigten sich (wenn auch geringe) Testleitereffekte aufgrund des Geschlechts in einer großen, repräsentativen Stichprobe von deutschen Probanden mit unterschiedlichem ethnischen und Bildungshintergrund.

Zusammenfassend zeigten die Ergebnisse der vier Studien, dass Testleiter die Testpersonen selbst bei standardisierten Tests beeinflussen. Erstens schienen sowohl Selbsteinschätzung und Erwartungen an die eigene Leistung, als auch die tatsächliche Leistung vom Testleitergeschlecht beeinflusst zu werden. Zweitens deuteten die Ergebnisse darauf hin, dass stereotypische Zuschreibungen zu verschiedenen Erwartungen an die Testsituation führten. In der Dissertation wird eine zusammenführende Diskussion der vier Studien präsentiert, in der praktische Implikationen und Forderungen an zukünftige Forschung beschrieben werden.

**Referenzen:**

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River (NJ): Prentice Hall.

Fernández-Ballesteros, R. (1999). Psychological assessment: Future challenges and progresses. *European Psychologist, 4*, 248-262. doi: 10.1027//1016-9040.4.4.248

Huang, M. H. (2009). Race of the interviewer and the black-white test score gap. *Social Science Research, 38*, 31-40. doi: 10.1016/j.ssresearch.2008.07.004

Mishra, S. P. (1980). Influence of Examiners Ethnic Attributes on Intelligence-Test Scores. *Psychology in the Schools, 17*, 117-122. doi: 10.1002/1520-6807(198001)17:1<117::AID-PITS2310170122>3.0.CO;2-6

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research*. New York, N.Y.: Irvington Publishers, Inc.

Rosenthal, R. (1995). Critiquing Pygmalion: a 25-Year Perspective. *Current Directions in Psychological Science, 4*, 171-172. doi: 10.1111/1467-8721.ep10772607

Sattler, J. M., & Theye, F. (1967). Procedural, Situational, and Interpersonal Variables in Individual Intelligence Testing. *Psychological Bulletin, 68*, 347-360. doi: 10.1037/h0025153

Socio-Economic Panel (SOEP) data for the years 1984-2009, version 26. (2010).

# Chapter 1

# Introduction

## Introduction

Psychological assessment procedures have become an inherent part in counseling, training, selection, and evaluation in many countries, not least in Europe (Fernández-Ballesteros, 1999; Nevo & Jäger, 1986): Especially the use of tests and standardized questionnaires has increased relevance (Meyer et al., 2003; Muniz et al., 2001; Oakland, 1997). Most people undergo several tests and examinations throughout their lives – for example, in the context of school, university, or work. Whereas ongoing technical developments, for example concerning item and test design, improve the applicability and informative value of tests (e.g., Daniel, 1997; Fahrenberg, 2001; Glas & van der Linden, 2000; Glas & van der Linden, 2003; Naglieri et al., 2004), a possible source of unsystematic variance remains less explored and controlled: the assessment setting and test situation.

Assessment settings may vary considerably in type and method: For example, tests may be presented as paper-pencil or computerized assessment, the setting may cover an individual examinee or a group, answers may be given orally or in written form (e.g., Anastasi & Urbina, 1997). Although tests themselves may further differ, they share the standardized process of evaluation and scoring of test taker responses (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). This thesis focuses on standardized performance tests that are applied in a face-to-face setting.

In most face-to face assessment situations two interacting parts can be differentiated: First, an examinee or test taker who is requested to show a certain behavior. Second, a person who administers the assessment instrument, commonly called *examiner* in different assessment contexts – like educational, vocational, or clinical assessment – and with reference to the use of different instruments – like tests, questionnaires, check lists (e.g., Anastasi & Urbina, 1997; Domino & Domino, 2006). With focus on standardized tests, more specific labels like *tester* or *test administrator* are used synonymously as well. All these terms apply to the administration of a psychological instrument and differentiate this part of the assessment process from scoring or interpretation of results. In accordance with the widespread use of examiner in the literature (e.g., Aiken & Groth-Marnat, 2006; Anastasi & Urbina, 1997; Domino & Domino, 2006), I will employ this term in the following thesis as well. The examiner is supposed to conduct the testing, establish a focused, positive working atmosphere, give information to the test taker, and impair cheating. Usually, the interaction during standardized testing is intended to aim for a professional, respectful, and attentive

setting. Therefore, examiners regularly receive prior training and are trained to give standardized instructions to the test taker.

Within the individual face-to-face testing the situation may also be described as a kind of social interaction between test taker and examiner. This situation can be influenced by attributes, personal beliefs, and opinions of the involved individuals (cf. Cronbach, 1956; Fiske & Taylor, 1996). Such a situation provides different demands for both sides: In many selection and admission tests the test taker is requested to perform in a mostly unknown test in a mostly unfamiliar setting. Following, the test taker has to handle a more or less socially demanding situation while aiming to show a maximum performance at the same time. For the examiner consequences of the testing are mostly hardly self-relevant. Furthermore, examiners' behavior is more predetermined due to the specified instructions and conventions (Argyle, 2009; Cronbach, 1956; Cronbach, 1984; Spitznagel, 1982). However, the risk of undesired examiner's influence on the test taker may be a substantial issue in psychological testing. This may be especially the case if due to a lack of knowledge regarding standardized testing procedures or based on prior test experience – e.g. in school exams – test takers ascribe examiners some power or control with reference to the results and the consequences of the testing. Therefore, the examiner may be perceived in an influential status raising test taker's sensitivity and susceptibility towards the examiner.

Primary goal of this thesis is to investigate if test takers' behavior respectively performance on standardized tests may be influenced by characteristics of the person who administers a standardized achievement test. Different biosocial person characteristics could be of relevance. In this thesis special regard is given to the gender as an inherent examiner characteristic.

This introduction will first outline the testing situation investigated in this thesis. Second, an overview on examiner influence from different research perspectives is given. Third, a conclusion on these investigations is proposed and open questions are raised. Subsequently, the potential relevance of examiner influence for the psychological testing practice is described. Then the major objectives of this thesis referring to the conclusions drawn from prior research are proposed. Finally, an overview of the four studies of this thesis and the upcoming chapters is given.

## 1.1 The Test Situation and Examiner Effects

The face-to-face test situation has been characterized as necessarily including elements of a social interaction (Brauns, 1982). Addressing this social interaction between one

examiner and one test taker, different basic sources of examiner differences could be distinguished possibly influencing the course of testing: First, examiner's conscious or automatic attitudes, motivations, and/or expectations could impact aspects of his or her administrating behavior. This could possibly include verbal behavior in unstandardized testing procedures. Moreover and irrespective from the content of the information given, the way to talk to the test taker – including voice, intonation, talking speed, or accent – is more or less individual and cannot easily be standardized in testing. Apart from such paraverbal characteristics, visible nonverbal behavior may also be a source of influence, as for example, examiner's expressiveness, mimics, or gait. However, some of these verbal and nonverbal behaviors could also be more standardized as a consequence of a very strict and consequent training. With reference to effects found in experiments, Rosenthal (1976) distinguished so called *active* and *passive* effects. Active effects arise from such described behavioral differences of the experimenters. Passive effects refer to the perception of experimenter characteristics that evoke a different performance of the participant. Therefore, even if the examiner was perfectly trained in verbal, vocal, and nonverbal behavior, there would still be biosocial characteristics left that may influence the testing procedure. These include demographic characteristics like gender, age, ethnicity, but also nonbehavioral aspects of appearance, as for example, persons' attractiveness. However, the performance of the test taker rarely allows for identification of either active or passive effects isolated.

Besides particular attributes related to such characteristics, as for example the pitch of the voice, there may be effects that origin in the associations and perceptions examiners evoke. As a consequence, examiners with different attributes may not only influence test takers' associations and perceptions, but may also lead to a change in interpreting the whole situation. This possible difference in interpretation of the test situation may, in turn, affect test takers' motivation, attention, well-being, and behavior. Not all of these person characteristics were investigated with reference to examiner effects. In following sections, I will give a brief overview of different research traditions investigating the possible influence of conductors – either experimenters or examiners- on research participants' behavior and test results. Early observations of such an influence were made in experimental research and occurred more or less accidentally. Later, in Psychological Assessment and test use the possibility of undesired influence was considered as well. In recent years, the investigation of unintended influences on test takers has gained interest in the context of fairness issues.

### 1.1.1 Early Findings on Experimenter Effects

The research on expectancy effects in terms of influencing the behavior of one subject by the own expectations and assumptions has a long history. One early, famous example did not even address humans and is that of "Clever Hans" described by Pfungst (1907). This horse was supposed to be able to solve mathematical tasks, spell, and read. However, Pfungst and others revealed that the horse was able to do so because it interpreted unintentional, subtle, nonverbal cues of his questioners and audience. Rosenthal (1994) came across this phenomenon when it appeared to him that he as an experimenter accidentally had influenced the subjects of his doctoral thesis according to his hypotheses in the 1950ies. From this time on Rosenthal (1976) investigated the influence of experimenters on participant's behavior thoroughly. On basis of his own work and reviews of prior research he distinguished different forms of experimenter influence: First, the experimenter's own motivations or expectations concerning the outcome may inadvertently be a source of influence. Such *experimenter expectancy effects* were shown to influence the learning success of rats (e.g., Rosenthal, 2002). Furthermore, a study by Rosenthal and Jacobson (1966) revealed that school children from whom teachers were told to expect high intellectual potential received higher results on an intelligence test after eight months. Expectancy effects were repeatedly shown in different settings – for example achievement and ability tests (Rosenthal & Rubin, 1978). Nevertheless, such effects on intelligence tests were not as robust (cf. Raudenbush, 1984). Over the last forty years experimenter expectancy effects revealed inconsistent results and were criticized for misinterpretation of data (e.g., Snow, 1995; Spitz, 1999).

Rosenthal (1976) subsumed a second group of influences due to the experimenter person under *experimenter effects* and found biosocial – e.g., gender or ethnicity – and psychosocial – e.g., dominance or anxiety – characteristics of the experimenter to be of relevance. Rosenthal (1976) reported differences due to experimenter gender in studies on motor performance, verbal learning, or picture rating tasks.

### 1.1.2 Early Findings on Examiner Effects

Referring to prior research, experimenter expectancy effects and experimenter effects were investigated to a large extend in experimental settings. One could propose that this does not fully refer to standardized test situations, as these are mostly defined by more standardized procedures, less variation of information provided, and more conformity in examiners' behavior – independent of the hidden expectations or beliefs. Therefore, experimenter expectancy effects may be reduced due to standardization (cf. Sattler & Theye, 1967).

Furthermore, psychosocial characteristics of the individual examiner should not become predominant in a standardized interaction. Nevertheless, biosocial characteristics can hardly be concealed. Therefore, the possibility of effects due to examiners' biosocial attributes remains for standardized test situations. The impact of such visible biosocial examiner characteristics on test takers' performance on standardized tests has been shown for ethnicity (Huang, 2009; Katz, Roberts, & Robinson, 1965; Mishra, 1980), age (Rosenthal, 1976), and attractiveness (Karremans, Verwijmeren, Pronk, & Reitsma, 2009). With regard to gender, a study by Samuel (1977) reported school children performing better when tested by a woman on an intelligence test. However, only one Black and one White female examiner were included. Therefore, individual differences cannot be ruled out. Cieutat (1965) found female examiners eliciting better results of children on an intelligence test.

### 1.1.3 Social Psychological Perspective on Examiner Effects

Besides investigation of the existence or generalizability of examiner effects, the background and mechanisms of such effects could also be addressed from a social psychological perspective. In such a social testing interaction two more or less unacquainted individuals are confronted and at least for the test taker the setting is unfamiliar and demanding. Therefore, it could be proposed that social categories are of major importance for perception and social judgments here (Allport, 1954). Stereotypes are social categories that comprise beliefs and personal theories about members of social groups (Hilton & von Hippel, 1996). The relevance of stereotypes for first impressions was described and investigated in different models (Brewer, 1988; Fiske & Neuberg, 1990; Kunda & Thagard, 1996). In Western societies gender and age are amongst the strongest physical cues for social categories (Fiske, 1993; Montepare & Zebrowitz, 1998; Tesser, 1988). Despite societal changes leading to women gaining more powerful positions in many countries (European Commission, 2011; U.S. Bureau of Labor Statistics, 2010), a study by Spence and Buckner (2000) showed gender stereotypes remaining stable over several decades. Stereotypical ascriptions towards typical men and women were mainly consistent between the 1970ies and the 1990ies: Women were perceived as more emotional, understanding, and compassionate, whereas men were perceived as more competitive, forceful, and aggressive. Furthermore, gender stereotypes seem to be influential for other person attributes as well: For example, research on voice perception recently revealed that masculinity was related to competence, independent of actual gender (Ko, Judd, & Stapel, 2009) and sexually dressed women were perceived as less

competent in mock application procedures (Glick, Larsen, Johnson, & Brandstiter, 2005; Wookey, Graves, & Butler, 2009).

Another social cue of relevance in first impressions is age (Fiske, 1993; Montepare & Zebrowitz, 1998). Data gained in previous studies indicated that apart from phenomena like ageism in the work context leading to derogation and devaluing of older workers (cf. Kite, Stockdale, Whitley, & Johnson, 2005; Perry, Kulik, & Bourhis, 1996; Posthuma & Campion, 2009), older employees were perceived as generally reliable, conscientious, and effective (Redman & Snape, 2002; Warr & Pennington, 1994).

### 1.1.4 Comment on Prior Research on Experimenter and Examiner Effects

Referring to preceding studies on experimenter influence and examiner effects in testing, they generally support effects of examiner's resp. experimenter's biosocial characteristics on test takers' behavior. Prior research on examiner's gender has been criticized for some drawbacks: Rumenik, Capasso, and Hendrick (1977) revealed no clear effect in their review. For adult participants their results indicated a small trend to male examiners eliciting better performance in achievement related tasks, whereas in some studies children performed better on an intelligence test when tested by a woman. However, the authors concluded that results were mixed and no clear conclusions for standardized test procedures could be drawn. The authors also argued that instruments employed varied considerably, impairing comparison of research results. Furthermore, the majority of studies were based on only few different examiners – some even only one man versus one woman.

Additionally, I would like to add several concerns: First, the earlier experimental studies addressed research questions different from test and assessment procedures. Therefore, the primary focus has not been laid on standardized test settings, test instructions, and standardized questions – or at least seldom standardized assessment conditions were reported. Most of the reported studies do not allow for conclusions with reference to standardization.

Second as already noted by other researches, studies using psychological performance tests rarely considered several examiners of different groups (e.g., Rumenik et al., 1977; Sattler & Theye, 1967). Therefore, systematic investigations of specific examiner characteristics as sources of influence are impaired by small sample sizes of examiners.

Third, apart from the investigation if the administrating person may influence the test takers, few propositions concerning potential underlying processes were made. In the context of experimental research, Rosenthal (1976) endorsed the notion of differential behaviors of

experimenters. In standardized test settings, where behavioral difference may be restricted, the perception of examiners may be more relevant.

Following, this thesis aims to further investigate the influence of examiner gender on test taker performance in standardized achievement tests. To overcome limitations of prior research with reference to its explanatory power in the domain of Psychological Assessment, special regard will be given to sample size – including both examiners and test takers – and application of contemporary test instruments.

## 1.2 Potential Relevance of Examiner Effects for Psychological Test Practice

In the context of Psychological Assessment, efforts were made to establish standardized, fair, and valid assessment procedures to provide valid and comparable results (cf. Standards for educational and psychological testing, AERA, APA, & NMCE, 1999; Guidelines for the Assessment Process, Fernández-Ballesteros et al., 2001; European Meta-Code of Ethics, Koene, 1997; DIN 33430, Westhoff et al., 2010). In 1954, the first elaborated standards and directions concerning psychological tests were published by the American Psychological Association (APA, cf. Novick, 1981): The *Technical Recommendations for Psychological Tests and Diagnostic Techniques* focused on issues like reliability, validity, administration, and norms. Since 1966 the organizations APA, AERA und NCME (cf. Novick, 1981) publish the *Standards for Educational and Psychological Testing.* The aim of these standards is to support the responsible use of tests and assessment instruments by providing basic criteria concerning the quality of tests and other diagnostic instruments. This section will first introduce two basic claims in psychological testing that may be affected by examiner effects. Second, a comment on the relevance of examiner effects concerning these two claims is presented.

### 1.2.1 Examiner Effects and Standards in Psychological Testing

Two of these basic claims with reference to the quality of assessment address its *objectivity* and *fairness*. The first, objectivity, refers to the fact that results are independent of the test setting and the examiner (Stuart-Hamilton, 1996; Westmeyer, 2003). Usually, objectivity refers to the person(s) behavior who administer(s) the testing. He or she is responsible for scoring, and interprets the outcomes. Therefore the examiner has to be regarded as being part of the test situation. Often the impression is made that objectivity can be obtained actively by training or adhering to standards described in a test manual. However, more generally, objectivity may be understood as a claim concerning the whole test situation

and process. Although objectivity may be regarded as predisposition for reliable, valid, and fair testing, possibilities to ensure objectivity in psychological testing are rarely outlined. Standardization is a common way to raise objectivity: During test administration standardization concerns specified instructions that should be given. Additionally, test manuals yield information concerning the provided test time, requested working materials, and sometimes even the seating of test takers in groups. The potential relevance of the testing situation as a social situation is hardly considered. However, even if the examiner is trained for standardized behavior, the mere presence of an administrating person may activate certain perceptions and expectations concerning the testing within the test taker. Therefore objectivity may be at risk even if the examiner sticks to the instructions. In a broader sense, objectivity is given if test results are invariant to situational features (including examiners) of the testing, whereas objectivity may not be claimed if test takers have different chances because somehow test conditions are differing. Objectivity may therefore be violated if test results are systematically influenced by situational aspects – for example the examiner – independent of test taker's characteristics.

The second standard, fairness in psychological achievement testing refers to equal opportunities to show maximum performance for all test takers, irrespectively of individual membership to different groups (see APA, AERA, & NCME, 1999). Fairness is a regarded issue in psychological testing since the development of early intelligence and aptitude tests: In the beginning of the 20$^{th}$ century, Binet (cf. McNamara & Roever, 2006) found results on an intelligence test to be influenced by the socio-economic status of the children and in 1912 a study discussed ethnicity to be a moderator of test results (Weintrob & Weintrob, 1912). However, the consideration of bias or fairness issues was incorporated much later in test standards, namely in the *Standards for Educational and Psychological Testing* in 1974. With reference to test fairness, nowadays three forms of bias may be distinguished to affect test results (van de Vijver & Poortinga, 1997): First, construct bias refers to differences in the construct between groups, e.g. cultural groups. Second, method bias applies to factors extraneous to the measured construct affecting test taker groups variably. Third, item bias concerns specific items that are answered differently by test takers according to their group membership and not to the construct. Helms (2006) further developed the conceptualization of fairness in psychological testing: In her *individual-difference model* she considered that the social groups commonly used in fairness investigations – for example gender or ethnicity – are proxies for underlying psychological characteristics of test takers. Following, with focus on the individual level any psychological characteristic may constitute an individual test

situation affecting the fairness of the testing. Concluding, fairness may be claimed if test takers with different characteristics (e.g. referring to social groups) have comparable chances. On the other hand fairness may not be claimed if test takers systematically have different chances under the same test conditions and this can be traced back to their shared characteristics.

Subsuming, the common basic idea underlying objectivity *and* fairness is the claim that test results are comparable between different test takers, because as a precondition it is ensured that *all test takers* performed under the *same conditions*. However, in practice, both sources of undesired variance could also be mixed, as will be shown in the following example: Imagine, several examinees are tested by examiner A and the other group is tested by examiner B. In the first case, all female test takers receive poorer results than the male test takers. This would question the fairness of the instrument because differences could be explained by test takers' shared characteristic – namely gender. As a second example, all test takers tested by examiner A could also receive poorer results than all test takers tested by examiner B. In this case, objectivity would be questioned, as situational impact given by the examiner would affect the performance. Furthermore, a third case could occur: Only women tested by examiner A would show poorer results than the other test takers. This would concern a combined fairness and objectivity problem of the testing.

As already referred to, test taker groups may also be defined on base of psychological characteristics in a wider sense, for example, when considering persons possessing higher or lower levels of test anxiety (see Ortner & Caspers, 2011). Considering fairness claims in psychological testing, test takers' gender is an obvious often referred to characteristic (Hough, Oswald, & Ployhardt, 2001; McNamara & Roever, 2006). For this reason and due to the fact that gender is ubiquitous in social interaction, test taker's gender will be regarded in this thesis on examiner effects.

Due to the unequal situational status of test taker and examiner (Spitznagel, 1982), this thesis further takes test taker's individual *social dominance orientation* (SDO) as characteristic into account. SDO refers to the individual appraisal of social hierarchies and the support of hierarchy-enhancing strategies compared with hierarchy-attenuating strategies (Sidanius & Pratto, 2001). People with high SDO generally support conservative politics; they hold ethnic prejudices, and oppose strategies for more gender equality. A study by Danso and Esses (2001) included the concept of SDO in a study investigating ethnical aspects of examiner effects and found that White participants with high SDO showed increased performance when tested by a Black examiner compared with test takers low in SDO and test

takers tested by a White examiner. By now, there is no publication investigating the influence of SDO on the experienced test situation with reference to gender.

### 1.2.2 Comment on the Potential Relevance of Examiner Effects

Surprisingly, there are currently no investigations of possible effects due to examiner gender for modern standardized testing. Given the impact decisions based on test results have – for institutions as well as for individuals – current results are needed. The question remains if current assessment methods, especially standardized tests, are prone to examiner effects. Whereas recently studies were conducted to investigate the impact of examiner ethnicity on test takers (Danso & Esses, 2001; Huang, 2009; Marx & Goff, 2005), there is a lack of research investigating examiner effects due to gender. If the claims of objectivity and/or fairness are not met in standardized testing the consequences of the testing have to be considered to be distorted.

## 1.3 Aims of this Thesis

This thesis was planned to investigate "situational influences" as effects arising from the administrating person on behavior respectively performance in standardized tests.

According to the last chapters, several research goals for this thesis are derived to investigate examiner effects in standardized performance testing. Limitations of prior research should be overcome and information on possible attributional processes contributing to examiner effects in face-to-face testing obtained. The main research objectives of this thesis can be described as follows:

1. Investigation if examiner gender influence test takers in standardized face-to-face testing.

Influence due to the examiners as part of the test situation would violate basic standards of psychological testing. First of all, the test situations' objectivity would be impaired. Prior research revealed contradictory results of examiner effects due to gender. Thus, current investigations of examiner effects in standardized test situations with special regard to gender are needed. This research is requested to use standardized tests, employ several male and female examiners, and consider the impact of stereotypes in social interaction to enlighten the emergence of examiner effects.

2. Exploration if examiners are perceived and evaluated in line with gender stereotypes.

With regard to the emergence of examiner effects the perception of examiners may be of importance. Rumenik et al. (1977) concluded that no positive ramification for either male or female examiners could be found. Their review was presented more than 30 years ago. Therefore we cannot conclude if either new tests or societal developments have led to changes in examiner effects due to gender. Indeed, for many women the vocational situation and occupational career options have improved leading to more women in powerful positions (cf. Diekman & Eagly, 2000; Eagly, 2003; European Commission, 2011). This development may support the assumption that nowadays gender should not matter in a test context. Nevertheless, social perception relies heavily on stereotypes (Fiske, 1993) that – once established – change slowly (Hilton & von Hippel, 1996). From a social psychological perspective existing stereotypes may shape the perception and evaluation of examiners. By now, there is no investigation of perception and evaluation of different examiners. Implications of different perceptions would arise for examiners – who may need training concerning behavior and appearance during administration – as well as for test takers – whose concentration and motivation to perform may depend on specific examiner characteristics.

3. Investigation if test takers – given the possibility to choose – show a preference for examiners differing in gender.

Usually, in psychological test procedures test takers are not given the choice upon who will administer the testing. Still, test takers' preferences for specific examiners may provide insight into possible expectations underlying examiners with different gender. With regard to test fairness it may be considered that examiner' gender may be of relevance only for some test takers. Due to the lack of results on this topic it seems important to explore the pattern of choice of test takers. Furthermore, it is an everyday observation at schools or universities that test takers often indicate preferences concerning the examiner. By now, there is a lack of research considering differential preferences in the context of test administration.

4. Analysis of the contribution of test taker characteristics to the interpersonal aspects in standardized testing.

As described in the introduction, the possibility that effects address only some test takers or address some test takers stronger in a performance test has to be considered. The evidence of such group based examiner effects would reduce objectivity and fairness of the

testing procedure. The test taker differences may not only pertain to behavior or performance, but also to perception and evaluation of examiners, and preference for specific examiners. Therefore the relevance of test taker characteristics to ascribed examiners' characteristics, preferences for examiners, or achievement under male or female examiners should be regarded. Test taker groups may be based on different psychological characteristics. In this thesis test taker's gender and test taker's SDO will be taken into account as relevant test taker characteristics in testing.

### 1.4 Overview

This thesis comprises four studies that were developed to add further insight in examiner effects on test takers in face-to-face test settings:

The first study (Chapter 2) was planned to investigate examiner effects due to gender in a face-to-face test setting. To overcome drawbacks of prior research ten female and ten male examiners participate in this study. Data from 93 nonpsychology students who participated voluntarily at the Free University Berlin was used. A subjective score – the self-estimated performance on a verbal knowledge test – and an objective score – the number of correctly solved items – was obtained by the chosen test. This study aimed to amend results to the first research objective, namely if examiner gender has an influence on test takers in face-to-face performance testing.

The second study (Chapter 3) aimed to add further insight in underlying processes of social interaction between test takers and examiners. Overall, 114 nonpsychology students were tested by one of 22 examiners. The method was similar to the second study to the extent that the same test was used in the same face-to-face setting at the Free University Berlin. In addition, perceived examiners' attractiveness was taken into account in this study. It was proposed that the individual perception of examiner attractiveness may influence the interaction between test taker and examiner. Although attractiveness seems to be an influential person characteristic (e.g. Dion, Berscheid, & Walster, 1972; Eagly, Ashmore, Makhijani, & Longo, 1991; Langlois et al., 2000), its impact on performance in assessment procedures is hardly investigated. This study was therefore planned to deepen the understanding of social perception and attractiveness in standardized testing. Therefore, the results should contribute to the first and second research goals: This study investigated, if the examiner gender influences the test taker's self-estimation and performance and if ascribed attractiveness of the examiner is involved in examiner gender effects.

The third study (Chapter 4) aimed to directly investigate the perception of examiners and to explore test takers' preferences for examiners. For the latter concern a pilot study was conducted where participants were requested to decide if they want to be tested by a specific male or female examiner in a real life setting. Although examiner effects are discussed for decades, currently no research exists on how test takers perceive examiners. An online survey with short video clips of several male and female examiners of two age groups giving a standard test instruction was developed to fill this gap. Possible differences among test takers were considered with regard to test taker gender and SDO. This study referred to the second, third, and fourth research objectives. More specifically the study investigated if examiners were perceived differently in accordance with existing gender stereotypes, if test taker's pattern of choice revealed preferences for male or female examiners, and if preferences of the test takers varied due to test taker characteristics.

The fourth study (Chapter 5) analyzed data from a large, representative survey of German households. The aim of this study was to overcome possible limitations of the first and second study using test data of only small sample sizes. It is the first study to date, that investigated effects of the examiner gender on participants' performances in a large survey. In 2006, participants of the German Socio-economic panel were invited to participate voluntarily in a short speed test. The test was applied in the individual household via laptop with the examiner being present during the testing. In the selected sample data of 2,863 participants tested by one of 178 examiners was analyzed. With multilevel analyses it is possible to investigate possible examiner gender effects while controlling for possible influences of individual examiners. This final investigation aimed to provide further knowledge concerning again the first research aim of this thesis, namely if examiner gender effects are revealed in face-to-face performance testing.

Finally, in Chapter 6 the results of these four studies are summarized and different conclusions proposed. The impact of stereotypical perception of examiners and the role of test taker characteristics are discussed. Implications for the test use are explicated. The chapter finishes with concluding remarks of this thesis.

## 1.5 References

Allport, F. H. (1954). The structuring of events: outline of a general theory with applications to psychology. *Psychological Review, 61*, 281-303.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing and Assessment* (12th ed.). Boston: Pearson Allyn & Bacon.

Argyle, M. (2009). *Social Interaction* (2nd ed.). Oxford England: Atherton Press.

Brauns, H.-P. (1982). Testen als Interaktion. Die Testsituation aus interaktionistischer und systemtheoretischer Sicht. Relationen zwischen Systemebenen. In: Hoefert, H.-W. (Ed.) *Person und Situation. Interaktionspsychologische Untersuchungen* (pp. 107-139). Hogrefe: Göttingen

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. J. Wyer (Eds.), *Advances in Social Cognition* (Vol. 1, pp. 1-36). Hillsdale, NJ: Erlbaum.

Cieutat, V. J. (1965). Examiner differences with the Stanford-Binet IQ. *Perceptual and Motor Skills, 20,* 317-318.

Cronbach, L. J. (1956). Assessment of individual differences. *Annual Review of Psychology, 7*, 173-196.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.

Daniel, M. H. (1997). Intelligence testing: Status and trends. *American Psychologist, 52*(10), 1038-1045. doi: 10.1037/0003-066X.52.10.1038

Danso, H. A., & Esses, V. M. (2001). Black experimenters and the intellectual test performance of White participants: The tables are turned. *Journal of Experimental Social Psychology, 37*, 158-165. doi: 10.1006/jesp.2000.1444

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin, 26*, 1171-1188. doi: 10.1177/0146167200262001

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*, 285-290. doi: 10.1037/h0033731

Domino, G., & Domino, M. L. (2006). *Psychological Testing: An Introduction* (2nd ed.): Cambrigde University Press.

Eagly, A. H. (2003). The Rise of Female Leaders. *Zeitschrift für Sozialpsychologie, 34*, 123-132. doi: 10.1024//0044-3514.34.3.123

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*, 109-128. doi: 10.1037/0033-2909.110.1.109

European Commission (Ed.). (2011). *Report on Progress on Equality between Women and Men in 2010 - The gender balance in business leadership*. Luxembourg: Publications Office of the European Union, 2011. doi: 10.2767/99441

Fahrenberg, J. (Ed.). (2001). Progress in Ambulatory Assessment: computer-assisted psychological and psychophysiological methods in monitoring and field studies. Seattle: Hogrefe & Huber

Fernández-Ballesteros, R. (1999). Psychological assessment: Future challenges and progresses. *European Psychologist, 4*, 248-262. doi: 10.1027//1016-9040.4.4.248

Fernández-Ballesteros, R., De Bruyn, E. E. J., Godoy, A., Hornke, L. F., Ter Laak, J., Vizcarro, C., et al. (2001). Guidelines for the Assessment Process (GAP): A proposal for discussion. *European Journal of Psychological Assessment, 17*, 187-200. doi: 10.1027//1015-5759.17.3.187

Fiske, S. T. (1993). Social Cognition and Social-Perception. *Annual Review of Psychology, 44*, 155-194. doi: 10.1146/annurev.ps.44.020193.001103

Fiske, S. T., & Neuberg, S. L. (1990). A Continuum of Impression-Formation, from Category-Based to Individuating Processes - Influences of Information and Motivation on Attention and Interpretation. *Advances in Experimental Social Psychology, 23*, 1-74.

Fiske, S. T., & Taylor, S. E. (1996). *Social Cognition* (2nd ed.). New York: McGraw Hill.

Glas, C. A. W., & van der Linden, W. J. (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer Academic Publishers

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*, 247-261. doi: 10.1177/0146621603027004001

Glick, P., Larsen, S., Johnson, C., & Branstiter, H. (2005). Evaluations of sexy women in low- and high-status jobs. *Psychology of Women Quarterly, 29*, 389-395. doi: 10.1111/j.1471-6402.2005.00238.x

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial group assessment: A quantitative perspective. *American Psychologist, 61*, 845-859. doi: 10.1037/0003-066X.61.8.845

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47*, 237-271. doi: 10.1146/annurev.psych.47.1.237

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152-194. doi: 10.1111/1468-2389.00171

Huang, M.-H. (2009). Race of the interviewer and the black-white test score gap. *Social Science Research, 38*, 29-38. doi: 10.1016/j.ssresearch.2008.07.004

Karremans, J. C., Verwijmeren, T., Pronk, T. M., & Reitsma, M. (2009). Interacting with women can impair men's cognitive functioning. *Journal of Experimental Social Psychology, 45*, 1041-1044. doi: 10.1016/j.jesp.2009.05.004

Katz, I., Roberts, S. O., & Robinson, J. M. (1965). Effects of task difficulty, race of administrator, and instructions on digit-symbol performance of Negroes. *Journal of Personality and Social Psychology, 2*, 53-59. doi: 10.1037/h0022080

Kite, M. E., Stockdale, G. D., Whitley, B. E., & Johnson, B. T. (2005). Attitudes toward younger and older adults: An updated meta-analytic review. *Journal of Social Issues, 61*, 241-266. doi: 10.1111/j.1540-4560.2005.00404.x

Ko, S. J., Judd, C. M., & Stapel, D. A. (2009). Stereotyping Based on Voice in the Presence of Individuating Information: Vocal Femininity Affects Perceived Competence but Not Warmth. *Personality and Social Psychology Bulletin, 35*, 198-211. doi: 10.1177/0146167208326477

Koene, C. J. (1997). Tests and professional ethics and values in European psychologists. *European Journal of Psychological Assessment, 13*, 219-228. doi: 10.1027/1015-5759.13.3.219

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*, 284-308. doi: 10.1037/0033-295X.103.2.284

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*, 390-423. doi: 10.1037/0033-2909.126.3.390

Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology, 44*, 645-657. doi: 10.1348/014466604X17948

McNamara, T., & Roever, C. (2006). Psychometric Approaches to Fairness: Bias and DIF. *Language Learning, 56* (Suppl 2), 81-128. doi: 10.1111/j.1467-9922.2006.00381.x

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2003). Psychological testing and psychological assessment: A review of evidence and issues *Methodological issues & strategies in clinical research* (3rd ed.). (pp. 265-345). Washington, DC US: American Psychological Association.

Mishra, S. P. (1980). The influence of examiners' ethnic attributes on intelligence test scores. *Psychology in the Schools, 17*, 117-122. doi: 10.1002/1520-6807(198001)17:1<117::AID-PITS2310170122>3.0.CO;2-6

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. *Advances in Experimental Social Psychology, 30*, 93-161.

Muniz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., et al. (2001). Testing practices in European countries. *European Journal of Psychological Assessment, 17*, 201-211. doi: 10.1027//1015-5759.17.3.201

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., et al. (2004). Psychological Testing on the Internet: New Problems, Old Issues. *American Psychologist, 59*, 150-162. doi: 10.1037/0003-066X.59.3.150

Nevo, B., & Jäger, R. S. (Eds.). (1986). *Psychological Testing: The Examinee Perspective*. Göttingen: Hogrefe.

Novick, M. R. (1981). Federal guidelines and professional standards. *American Psychologist, 36*(10), 1035-1046. doi: 10.1037/0003-066X.36.10.1035

Oakland, T. (1997). Test use among school psychologists: Past, current, and emerging practices. *European Journal of Psychological Assessment, 13*, 2-9. doi: 10.1027/1015-5759.13.1.2

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment, 27*, 157-163. doi: 10.1027/1015-5759/a000062

Perry, E. L., Kulik, C. T., & Bourhis, A. C. (1996). Moderating effects of personal and contextual factors in age discrimination. *Journal of Applied Psychology, 81*, 628-647. doi: 10.1037/0021-9010.81.6.628

Pfungst, O. (1907). Das Pferd des Herrn von Osten. Der Kluge Hans: Ein Beitrag zur experimentellen Tier- und Menschenpsychologie. Leipzig: Johann Ambrosius Barth.

Posthuma, R. A., & Campion, M. A. (2009). Age Stereotypes in the Workplace: Common Stereotypes, Moderators, and Future Research Directions. *Journal of Management, 35*, 158-188. doi: 10.1177/0149206308318617

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76*, 85-97. doi: 10.1037/0022-0663.76.1.85

Redman, T., & Snape, E. (2002). Ageism in teaching: stereotypical beliefs and discriminatory attitudes towards the over-50s. *Work, Employment and Society, 16*, 355-371. doi: 10.1177/095001702400426884

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research* (enlarged ed.). New York, N.Y.: Irvington Publishers, Inc.

Rosenthal, R. (1994). Interpersonal Expectancy Effects: A 30-Year Perspective. *Current Directions in Psychological Science, 3,* 176-179. doi: 10.1111/1467-8721.ep10770698

Rosenthal, R. (2002). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist, 57*, 839-849, doi: 10.1037/0003-066X.57.11.839

Rosenthal, R., & Jacobson, E. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports, 19*, 115-118.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 1*, 377-415. doi: 10.1017/S0140525X00075506

Rumenik, D. K., Capasso, D. R., & Hendrick, C. (1977). Experimenter sex effects in behavioral research. *Psychological Bulletin, 84*, 852-877. doi: 10.1037/0033-2909.84.5.852

Samuel, W. (1977). Observed Iq as a Function of Test Atmosphere, Tester Expectation, and Race of Tester - Replication for Female Subjects. *Journal of Educational Psychology, 69*, 593-604. doi: 10.1037/0022-0663.69.5.593

Sattler, J. M., & Theye, F. (1967). Procedural, situational, and interpersonal variables in individual intelligence testing. *Psychological Bulletin, 68*, 347-360. doi: 10.1037/h0025153

Sidanius, J., & Pratto, F. (2001). *Social Dominance* Cambridge: University Press.

Snow, R. E. (1995). Pygmalion and intelligence? *Current Directions in Psychological Science, 4*, 169-171. doi: 10.1111/1467-8721.ep10772605

Socio-Economic Panel (SOEP) data for the years 1984-2009, version 26. (2010).

Spence, J. T., & Buckner, C. E. (2000). Instrumental and expressive traits, trait stereotypes, and sexist attitudes. *Psychology of Women Quarterly, 24*, 44-62. doi: 10.1111/j.1471-6402.2000.tb01021.x

Spitz, H. (1999). Beleaguered Pygmalion: A history of the controversy over claims that teacher expectancy raises intelligence. *Intelligence, 27,* 199-234. doi: 10.1016/S0160-2896(99)00026-4

Spitznagel, A. (1982). Die diagnostische Situation. In C. F. Graumann, H. Feger & J. Bredenkamp (Eds.), *Enzyklopädie der Psychologie. Themenbereich B Methodologie und Methoden, Serie II. Psychologische Diagnostik* (Vol. 1, pp. 248-294). Göttingen: Hogrefe.

Stuart-Hamilton, I. (1996). Dictionary of psychological testing, assessment, and treatment (Rev. ed.). London:Kingsley

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 181-227). New York: Academic Press.

U.S. Bureau of Labor & Statistics (Ed.). (2010). *Highlights of Women's Earnings 2009*.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment, 13, 29-37. doi: 10.1027/1015-5759.13.1.29

Warr, P., & Pennington, J. (1994). Occupational age-grading: Jobs for older and younger nonmanagerial employees. *Journal of Vocational Behavior, 45*, 328-346. doi: 10.1006/jvbe.1994.1039

Weintrob, J., & Weintrob, R. (1912). The influence of environment on mental ability as shown by Binet-Simon Tests. *Journal of Educational Psychology, 3*, 577-583. doi: 10.1037/h0071857

Westhoff, K., Hagemeister, C., Kersting, M., Lang, F., Moosbrugger, H., Reimann, G., et al. (2010). *Grundwissen für die berufsbezogene Eignungsberurteilung nach DIN 33430* (3rd ed.). Lengerich: Pabst Publisher.

Westmeyer, H. (2003). Objectivity. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 629-632). London: Sage.

Wookey, M. L., Graves, N. A., & Butler, J. C. (2009). Effects of a sexy appearance on perceived competence of women. *Journal of Social Psychology, 149*, 116-118. doi: 10.3200/SOCP.149.1.116-118

# Chapter 2

## Test Administrator's Gender Affects Female and Male Students' Self-estimated Verbal General Knowledge

Ortner, T. M. & Vormittag, I. (2011). Test administrator's gender affects female and male students' self-estimated verbal general knowledge. *Learning and Instruction, 21,* 14-21. Published version available at: http://dx.doi.org/10.1016/j.learninstruc.2009.09.003

## 2.1 Abstract

Effects of test administrator's gender on test takers' self-estimated verbal general knowledge and de facto verbal general knowledge were investigated. Based on three theories previously applied in research dealing with the effects of test administrator's ethnicity, it was expected male and female test takers to show higher scores under female test administration. In a double-blind face-to-face-testing design, 93 university students of both genders in four groups were tested by 20 test administrators of both genders. A MANOVA confirmed the expected significant main effect. Female and male students reached higher scores in self-estimated knowledge when tested by a female test administrator in comparison to female students and male students tested by a male test administrator (Cohen's $d$ = 0.46). No significant effects resulted for de facto knowledge.

*Keywords:* Test administration effects; Gender stereotype; Stereotype threat; Self-estimation; Metacognition

# Test Administrator's Gender Affects Female and Male Students' Self-estimated Verbal General Knowledge

## 2.2 Introduction

Performing a cognitive task during a learning process, on an exam, or in a standardized test situation is a dynamic process that is always situated in a particular context. Acquired knowledge does not necessarily lead to a successful performance (Schutz and Davis 2000). Within the field of educational and employment testing, efforts are being made to reduce assessment bias by targeting effects that are believed to systematically impair test performance. A fair assessment procedure is supposed to provide comparable opportunities for examinees to demonstrate acquired knowledge and skills that are relevant to the test's purpose (Willingham and Cole 1997). However, this requirement is often not given in testing practice; even within standardized testing procedures, ethnic, socioeconomic, or gender characteristics might have an impact on performance. In recent times, several mechanisms and testing conditions have been identified as leading to systematic differences in test results – a form of bias (Wheeler and Petty 2001; Marx and Stapel 2005). Related to this issue is the objectivity of an assessment procedure, that is, the independence of test results from the testing situation and the test administrator; this also must be considered as a possible source of systematic differences in test results.

In past discussions, surprisingly little attention has been paid to the effects generated by the person who administers a testing procedure, especially with reference to gender. The test administrator, including his or her gender, cannot be purged: There is always at least one person who is responsible for introducing the test, for informing and securing consent in testing, as well as for intervening in cases of potentially bias-evoking conditions (e.g. Fernandez-Ballesteros, De Bruyn et al. 2001). This person is a situational characteristic of the test, and could possibly evoke bias. Hence, we need to know more about systematic effects related to test administrator's characteristics. The following manuscript addresses effects of test administrator's gender on the self-estimation and performance of men and women when taking a test of general knowledge.

### 2.2.1 Test Administrator's Gender Effects

Early studies have already tested the assumption that not only do the examinee's or interviewed person's abilities and characteristics lead to certain outcomes, but the conductor's characteristics and behaviours can also have an effect. Specifically, the effect of one person's

expectations on others' behaviours was part of several early experiments (Rosenthal and Fode 1963; Rosenthal and Jacobson 1968). It has been shown that manipulated expectancies of both examiner (teacher) and tested person (student) can interact and have systematic effects on intellectual performance (Zanna, Sheras et al. 1975; Raudenbush 1984). In addition to expectancy effects, several studies have also revealed an experimenter bias when experimentally manipulating the experimenter's behaviour, that is, persons taking a general aptitude test from a positive or neutral test administrator scored higher than persons tested by a negative examiner (Bookout and Hosford 1969).

Very few experiments have focused on possible gender effects. However, with reference to recent studies that focused on test administrator's ethnicity, three different effects can be transferred to the domain of gender for hypothesis forming. First, a more general theoretical framework from research on test bias has been recently applied to the domain of test administration effects, namely the concept of *stereotype threat*. Situations in which stereotypes negatively affect the target person may lead to stereotype threat, an impairment of task performance (Steele 1997; Spencer, Steele et al. 1999). It has been proposed that subgroup differences in test performance are caused by the harassment a person perceives as a consequence of a testing situation in which one may be at risk to confirm an existing negative stereotype (Steele and Aronson 1995). Research on underlying processes leading to stereotype-conforming performance under threat has revealed higher arousal, stress, and anxiety in target persons (Ben-Zeev, Fein, & Inzlicht, 2005; O'Brien & Crandall, 2003; Quinn & Spencer, 2001; Schmader, Johns, & Forbes, 2008; Wheeler, Jarvis, & Petty, 2001).

Marx and Goff (2005) applied the concept of stereotype threat to test administrator effects with reference to the test taker's race. In their experiment they investigated whether Black undergraduates would experience higher levels of threat when tested by a White experimenter than when tested by a Black experimenter. They further assumed that the presence of a Black experimenter would attenuate the effects of threat for performance on a verbal test. Indeed, results showed that Black participants with a Black test administrator (a) outperformed those with a White test administrator, and (b) also described feeling less threatened by the test-taking situation. From a gender perspective, negative stereotypes also exist with regard to the intellectual abilities of women (Beloff 1992; Reilly and Mulhern 1995; Bennett 1996). Therefore, results similar to those reported by Marx and Goff (2005) for women tested by a woman or a man, respectively, would be expected. Based on stereotype threat theory, no effects for men should be expected as there would not be any threat in any of the situations.

If no other standard has been declared in a testing situation, Katz, Roberts, and Robinson (1965) suggested that characteristics of the tester – here, the ethnicity – might also influence the examinee's expectations of a given test's difficulty. Specifically, Katz et al. (1965) assumed that Black students would expect White students to score higher on the given tests in general. Thus, for Black test takers, the presence of a White test administrator could imply that their results would be compared to a more difficult *frame of reference*, namely the reference of White students. Based also on their experiments at American universities in southern states, they explained the performances of Black students in the presence of a White test administrator – as compared to a Black test administrator – as an effect of motivation. Being tested by a White experimenter generally increased their motivation to succeed. However, in combination with an ethnically biased aptitude test, they observed an over-arousal and thus, lower scores compared to a testing situation with a Black experimenter. This effect also seems possible with respect to the test administrator's gender; for example, testing an ability affected by gender differences or stereotypes, male and female test administrators should have different effects on women's performance. If test administrator's gender is interpreted as a cue for the frame of reference given in the testing situation, and if it has an impact on motivation, this should be shown by lower performances for women tested by a man than tested by a woman, at least, if the assessed aptitude is supposed to be gender biased. In line with Katz et al. (1965), the motivation of male participants should not be influenced by the experimenter's gender. Therefore, according to this framework, no differences for male participants due to experimenter's gender should be expected.

A third framework for the interpretation of test administration effects has been proposed by Danso and Esses (2001). In their research, they assessed the performances of Black and White students tested by test administrators of the same or a different ethnic group. They referred to *social dominance* theory (Sidanius and Pratto 1999), which states that in groups of unequal power, ideologies are formed to justify and maintain group hierarchy. They supposed that the superior social status of Whites in American society – compared to Blacks – might make Whites believe that this position is legitimated because they are superior with reference to intellectual abilities. In the presence of a Black test administrator, this could cause an effort to maintain such a perception for White test takers. As another possible effect, it was stated that White test takers might show better performance in the presence of a Black test administrator because they feel especially sure of themselves when compared to Blacks. In fact, White participants who were tested by a Black experimenter showed better

performance on a test requiring arithmetic operations with self-reported social dominance orientation moderating the effect (Danso and Esses 2001).

How could social dominance theory serve to predict gender effects of test administrators? Although the situation of women and their social status has improved within the last century, women are still rarely found in major positions of public leadership (Carli and Eagly 2001). Furthermore, women lack social influence compared to men (Carli 2001); for example, in Germany in the year 2006, there were no women as chief executive or president of the country's 50 largest companies; only about 12% of the board of managing directors were women, and only 9% of high-ranking positions (professorships) at universities were held by women (Eurostat 2008). Therefore, gender groups can be seen as unequal power groups in Germany as well as in various other countries. It seems possible that, at least in some domains, men believe that they are superior, and refer to their intellectual abilities to legitimate their position. With regard to gender, social dominance orientation may also lead men to show more effort if tested by a woman, legitimating their status, or helping them to feel especially sure of themselves, and thus helping them to achieve better results. Hence, men tested by a woman would outperform men tested by a man.

What results can be expected for women tested by a woman (vs. tested by a man) with reference to social dominance orientation? Women generally report a lower social dominance orientation (Pratto, Stallworth et al. 1997) and lower competitiveness (e.g. Niederle and Versterlund 2007) than men. However, in a more recent study involving real behaviour samples, Schmid Mast (2002) analyzed interruptions in experimental discussions as a sign of dominance in all-male versus all-female groups; overall, she found significantly more interruptions in female than in male groups. The results were interpreted as competitiveness in all-female groups as well as a female tendency to strive for dominance in same sex groups. It has also been shown that women differ in their attribution of men's and women's success to good luck, that is, women tend to attribute the successes of other women more to good luck than the successes of male stimulus persons (Deaux and Emswiller 1974), especially if the women are perceived as attractive (Försterling, Preikschas et al. 2007). Considering the attractiveness of the status of a test administrator (or the status of research assistants in the current study), this may evoke competitiveness in women, and may result in female test-takers denying the higher abilities, and thus, the legitimation of female test administrators. Therefore, similar outcomes of dominance seem possible for women against women as was proposed for the men. Therefore, from a social dominance perspective, it was expected both men and women to perform better if tested by a female test administrator than by a male one.

Based on the already existing results found for people of different ethnicities, it can be assumed with reference to gender that men will perform better when tested by a woman than by a man examiner and women will perform better when tested by a female examiner compared to being tested by a male. This assumption is based on all three described theories, that is, (a) the stereotype threat hypothesis, (b) the expected frame of reference in the absence of other cues, and (c) the considerations about social dominance.

#### 2.2.1.1 Test Administrator's Gender Effects in Educational Context

There is at least one study on high school students that revealed better results for both genders when the students were tested by a female test administrator as compared to a male (Samuel 1977). However, this study was conducted on adolescents, and only used four different male and female test administrators, which might have increased individual differences effects between the test administrators. The results were also published 30 years ago. As the situation of women in society has changed over the last decades (BMFSFJ 2005), with a tendency toward emphasizing gender similarities more than differences (Hyde 2005), it would be interesting to investigate whether gender-related test administration effects exist now, and how they are shaped.

The question of administration effects is not only of interest for a better understanding of testing in general and educational testing in particular. Situations in which participants are writing knowledge exams or performing exercise tasks in a learning context share elements with standardized testing situations, that is, in both situations students are given the opportunity to demonstrate acquired knowledge and the outcome is evaluated with reference to given standards. The main differences between the learning and the testing situations are that (a) tests are mostly conducted by unfamiliar persons (as opposed to being given by familiar teachers), (b) the testing situation is highly standardized (especially, *what* to say), and (c) the testing situation is often perceived by test takers as an *ability-diagnostic* situation; emphasizing this potentially threatening aspect before testing has been shown to lead to task impairment (Steele and Aronson 1995). Psychological tests are therefore assumed to be better able to objectively "measure" one's abilities. By contrast, teacher's evaluations are interpreted as less objective, and are considered to be a form of *interpersonal evaluative feedback* (Jussim, Coleman et al. 1989). Negative evaluations are, for example, occasionally perceived as an indicator that teachers hold an inaccurate unfavourable impression of students (Coleman, Jussim et al. 1987). In sum, although performing tasks in learning situations and in test situations share common aspects, we expect the testing situation to be perceived as more

self-relevant and activating. Studies have shown stronger effects of stereotyping in situations that are perceived as a potential threat to self-esteem (Fein and Spencer 1997; Spencer, Fein et al. 1998). As all test administrator effects described above include stereotyping mechanisms, we expect such effects in particular for the potentially more self-relevant testing situation.

### 2.2.2 The present study

To test systematic effects of test administrator's gender, two types of measures as dependent variables were included in the present study, a *subjective* and an *objective* measure. First, a subjective score was built by the number of items a person estimates that she or he will be able to solve. It was employed as an indicator of prospective estimate of solution correctness which is a metacognitive experience (Efklides, Samara et al. 1999; Efklides 2006). Metacognitive experiences are defined as a person's thoughts, judgments/estimates**,** and feelings when coming across a task and processing task-related information (Efklides 2006; Efklides 2008). Emotions and metacognitive experiences are present throughout situations of learning or task performance, and are triggered by situational characteristics and the person's appraisals (Efklides, Samara et al. 1999; Efklides and Volet 2005). More than objective performance, metacognitive experiences can therefore provide insight into a person's inner states and feelings during task processing. Such experiences during task processing shape expectations and goals for the current but also for future learning processes (Efklides 2008).

Second, an *objective* score was employed. The question of whether objective measures are affected by test administrators' characteristics addresses problems of fairness if only a particular group of persons is impaired. Differences between groups would indicate problems of objectivity if persons under certain testing conditions, independent from their personal characteristics, would be impaired in task performance. Viewing the test situation as a dynamic process, we hypothesized that – with reference to test administrator's gender – the comparison of subjective scores will evoke greater effects than the comparison of objective scores, as metacognitive experiences do not directly result in objective performance scores. The latter are further influenced by skills and knowledge, as well as successful self-regulatory processes (Schutz and Davis 2000).

The research question was, therefore, whether the test administrator's gender has effects on the performances of female and male students on (a) the self-estimation of their verbal general knowledge before performing the task, and (b) a task assessing de facto verbal general knowledge. An experimental design with two independent variables and two

dependent variables was chosen. Test-takers' and test administrators' gender were the two independent variables. As dependent measures, we used a test providing two pieces of information: (a) a score for *self-estimation* of the person's own verbal general knowledge, and (b) a score for *de facto* verbal general knowledge. It was expected, under standardized conditions, the experience of having a female test administrator will lead to higher subjective scores and better objective scores when compared to the experience of having a male test administrator. Therefore, it was hypothesized that test takers would show higher self-estimations (subjective score) if tested by a woman than if tested by a man (Hypothesis 1). Second, it was hypothesized that test takers would show better de facto knowledge (objective score) if tested by a woman than if tested by a man (Hypothesis 2). Third, it was hypothesized that the comparison of subjective (self-estimation) scores would induce greater effects than the comparison of objective (de facto knowledge) scores (Hypothesis 3).

## 2.3 Method

### 2.3.1 Design

In an experimental approach, students were randomly assigned to one of four groups: female students tested by a woman (Group 1, $n$ = 32); female students tested by a man (Group 2; $n$ = 21); male students tested by a woman (Group 3, $n$ = 19), male students tested by a man (Group 4, $n$ = 21). Test-takers first worked on a test module for assessing their self-estimated verbal general knowledge; after this, they worked on a test assessing de facto verbal general knowledge.

### 2.3.2 Participants

Test takers were 93 (53 women aged 18-29 years, $M$ = 23.6, $SD$ = 2.7, and 40 men aged 18-30 years, $M$ = 24.5, $SD$ = 3.1) university students. They were approached in public places at the university and asked to participate by student assistants. They were tested in a quiet room at the department. They were informed in advance that a facet of intelligence would be tested. Participation in the testing was voluntary; psychology students were excluded from participation. The four groups of test takers did not differ significantly in age, $F(3, 89) = 1.44$, $p = .24$.

Test administrators were 20 advanced students in psychology who were blind to the aims and hypotheses. All test administrators were European (10 women aged 20-37 years, $M$

= 24.7, *SD* = 4.6, and 10 men aged 21-42 years, *M* = 29.2, *SD* = 7.0). They were recruited by a notice posted on campus, and received an expense allowance of 8 euros per hour of testing. Age difference between the men and women as test administrators was not significant, $t(14) = -1.01$, $p = .31$.

### 2.3.3 Materials

**Self-estimated and de facto verbal general knowledge.** For assessment of test takers' self-estimated and de facto verbal general knowledge, items out of the computerized Lexical Knowledge Test (LKT; Wagner-Menghin 2004) were applied. The LKT test battery includes two modules, namely self-estimation and de facto verbal general knowledge.

In the first module of the test, test takers are given word lists (10 words). They have to estimate (Yes/No) whether they know *and* are able to explain the words on the list, and they are informed that they will subsequently have to explain the words. This task was used to determine a score for self-estimation. The score for self-estimated verbal general knowledge was calculated by summing the words a test taker declared to know *and* to be able to explain (two points for each, maximum 20).

In the second module, test takers have to fill in two missing phrases in sentences by choosing one option from a list of three to six phrases for each. The LKT main module, therefore, assesses crystallized intelligence (gc; Horn and Cattell 1966). The 10 items used were chosen from a level of medium difficulty. The items represent different fields of verbal general knowledge, for example, art ("copper engraving"), medicine ("oligophrenia"), and nutrition ("calvados"). This task was used to determine a score for de facto verbal general knowledge. The score for de facto verbal general knowledge was calculated as the total number of correctly filled in phrases; therefore, a total score of 20 could be gained since there were two responses per item.

Internal consistency for the de facto verbal general knowledge was higher (Cronbach's $\alpha = .52$) than for self-estimated verbal general knowledge (Cronbach's $\alpha = .35$). Low internal consistencies were anticipated as recent studies indicated low intercorrelations between different domains of verbal general knowledge (see Lynn, Wilberg et al. 2004).

### 2.3.4 Procedure

All testings were conducted in a single face-to-face setting. Test takers were welcomed by the test administrator, and general information was given about the duration and content of the study ("This is a study about applicability of group norms in a face-to-face-testing"). The test administrator read standardized instructions. The testing started with the

self-estimation knowledge module ("Which of the following words do you know and are able to explain?"). After this, the de facto knowledge module task was presented ("Please tell me which of the response options is the correct completion for the description of the word."). To avoid an influence on the responses due to intonation by the test administrator, test takers were requested to read each item by themselves and to tell their choice to the test administrator (e.g., "I chose options b and f."). At the end, test takers were asked their age. Including instructions, every testing took about 10 -15 min.

### 2.3.5 Statistical analysis

For the calculation of group differences in self-estimated verbal general knowledge and de facto verbal general knowledge, a MANOVA was performed, using test takers' gender and test administrator's gender in the analyses as fixed factors, and the two test scores as dependent variables. We applied two-tailed significance tests in all statistics.

## 2.4 Results

Descriptive data for all four groups are given in Table 2.1. The 2(test taker gender) x 2(test administrator gender) MANOVA including the self-estimation and de facto knowledge scores as dependent variables revealed the following results. Specifically, a significant main effect of a test taker's own gender on the dependent variables was found, Wilks's $\lambda = 0.90$, $F(1, 91) = 5.11$, $p = .01$, partial $\eta^2 = .10$, with female test takers scoring lower than males, as well as a significant main effect of test administrator's gender, Wilks's $\lambda = 0.93$, $F(1, 91) = 3.33$, $p = .04$, partial $\eta^2 = .07$. Scores gained under female test administration were higher than scores under male test administration. No significant Test Taker Gender x Test Administrator Gender interaction was revealed, Wilks's $\lambda = 0.99$; $F(3, 89) = 0.60$, $p = .94$.

*Table 2.1.* Means (and *SD*) of test takers' performance on self-estimated verbal knowledge and de facto verbal knowledge modules as a function of test administrator and test taker genders

| | Test Administrator | | | |
|---|---|---|---|---|
| | Female | | Male | |
| Test takers | Male | Female | Male | Female |
| Self-estimated knowledge | 14.42 (2.80) | 13.06 (2.03) | 12.67 (2.92) | 11.52 (2.89) |
| De facto knowledge | 13.11 (2.26) | 11.69 (1.73) | 12.00 (2.15) | 11.05 (1.88) |

*Note*. Means and standard deviations of raw scores (possible values range from 0 to 20).

The univariate ANOVAs revealed a significant main effect of test administrator gender for self-estimated knowledge, $F(1, 91) = 5.99$, $p = .02$, but not for de facto knowledge, $F(1, 91) = 1.65$, $p = .20$. Again, higher self-estimations were found under female test administrators. Effect size was Cohen's $d = 0.46$ for differences in self-estimated knowledge under female versus male test administration. Referring to Cohen's (1992) frame of interpretation, the effect size indicates an almost medium effect.

## 2.5 Discussion

Testing effects of test administrator's gender confirmed Hypothesis 1 that test takers' self-estimated knowledge can be systematically affected by the gender of the person who is administering a standardized test. The results of the present study showed that both male and female test takers gave higher estimations of their knowledge (subjective score) when a woman administered the test compared to a man. However, Hypothesis 2 was not confirmed. Specifically, test takers did not demonstrate greater de facto knowledge (objective score) under female administration. Indeed, in Hypothesis 3 a smaller effect for the objective score than for the subjective score was expected.

The effect size of the test administrator effect for subjective scores between the groups reached almost medium level. Willingham and Cole (1997) emphasized, with reference to assessment procedures, that even very small group differences may produce great factual effects if only few persons are selected from a large population. Although we did not find differences in objective performance, the given result is relevant for situations of selection; this is the case, if persons are only invited to talk about their knowledge and their strengths,

for example, in an interview, and if they are not required to demonstrate it, and if only a small amount of persons is selected on base of this information.

The relevance is also not only given with reference to situations of testing: We expect similar, although less strong results for situations of writing exams and performing tasks in a learning situation. Effects are then supposed to be more confounded by particular characteristics of known instructors and supposed to be weaker, as those situations are mostly seen as less ability-diagnostic, less self-esteem threatening, and grades and feedback are seen more as *interpersonal evaluative feedback* (Jussim et. al.1989). Future studies should address metacognitive experiences as dependent upon instructor's gender rather than the person's knowledge base. However, we would assume that the effects also are present. In case of a male teacher, students may start with lower expectations for successful task completion and may also avoid help-seeking behaviour (Turner, Thorpe et al. 1997, March).

The given result cannot be interpreted under the light of the stereotype threat theory (Steele 1997), as both men *and* women had lower estimations under test administration by a man. However, from the given results, it can only be concluded that one or more of the theories presented in the theoretical introduction can explain the underlying processes. It must be taken into consideration that different effects might have impacted test takers with different characteristics in different ways in our experiment; for example, stereotype threat might explain the results of (some) women in our experiment, whereas social dominance theory (Sidanius and Pratto 1999) might explain the results of person's with a higher social dominance orientation. For another group of test takers, the anticipated frame of reference may have influenced their performance. According to stereotype threat theory, if tested by a female test administrator, female test takers will not experience harassment, and will not be at risk of confirming an existing negative stereotype, as opposed to being tested by a man. Having a female test administrator saved the women from anxiety, higher arousal, and stress, and therefore, they showed higher results under this condition (Wheeler, Jarvis et al. 2001; Wheeler and Petty 2001). Referring to social dominance theory and Danso's (Danso & Esses, 2001) conclusions from his ethnicity research, it is possible that in the testing situation where an academic domain is made salient, having a female test administrator implies that progress is being made (since the academic domain has previously been typically dominated by men). Thus, people with a high dominance orientation may have been especially motivated to perform well to prove their superiority, *and/or* may have been feeling especially sure about their success in this testing situation. According to the expected frame of reference, test difficulty may also have been estimated as lower in the presence of a female test

administrator, which may have been accompanied by lower feelings of anxiety and arousal as compared to the presence of a male test administrator. However, to prove the relevance of each interpretation, additional data are required.

Additionally, future studies on test administrator's effects should also explicitly address the role of the test taker's experienced *affect* on self-estimations and test results. Positive mood is known to facilitate adaptive self-regulation (as careful processing of goal selection and goal-relevant information) and to enable persons to overestimate their chances of attaining a good outcome (Aspinwall 1998). Higher subjective and objective scores might serve also as outcome of good mood induced by the female test administrators compared to the mood induced by male test administrators. Positive mood might, therefore, serve as a mediator between all three potential test administrator effects (stereotype threat, test administrator's gender as a cue for task difficulty, and social dominance), and self-estimated knowledge. Mood may play a key role, especially with reference to learning processes. Efklides and Petkaki (2005) showed effects of a mood treatment on metacognitive experiences in the domain of mathematics. Mood predicted interest, liking, and also feeling of difficulty. In line with our result, they also revealed no effect on objective maths performance. With reference to a learning situation, our result may indicate that an unknown female teacher or a female instructor might lead to more positive mood – at least at the very beginning – than a male teacher or instructor.

There are limitations in the present study which should be mentioned. Specifically, the present study was conducted in Germany. With reference to Hofstede (2001), Germany holds the rank 9/10 of 53 nations and regions with reference to the dimension of masculinity as an indicator for gender role distribution. Generalization of the results according to the unequal power of gender groups may not apply or apply less in countries which feature higher equality with reference to gender roles.

Second, the present study addressed self-estimated verbal general knowledge and de facto verbal general knowledge. Research showed that in most different domains of general knowledge men gain better results than women (Lynn, Wilberg et al. 2004). Whereas there are no studies that specifically address gender differences in self-estimated general knowledge, it was repeatedly shown that men tend to estimate their intellectual abilities and past performances higher than women (Rammstedt and Rammsayer 2000; Rammstedt and Rammsayer 2002; Sieverding 2003). A recent study investigated feelings of confidence with regard to solving a mathematical task after reading it (Boekaerts and Rozendaal, in press) and showed higher feelings of confidence for boys. It is of future interest to determine whether the

same effects would result if testing different cognitive abilities, gender-neutral abilities, and also, abilities that favour women.

Furthermore, additional studies should also focus on systematic differences in the behaviours of men and women as test administrators. Gender effects can possibly be erased by the examination of behaviour. For example, some early studies investigating experimenter effects suggested that men smiled less often than women (Rosenthal 1976). If differences in behaviour can be found, at least some of them might be reduced by a more elaborated test administrator training – or, again, computerized assessment.

Future studies should still investigate how this bias can be overcome. Possible solutions might be the presence of several test administrators of different genders during testing, or computerized presentation of tests. However, neither education nor educational testing can function fully without the presence of human beings; for example, some tests might not be amenable to being conducted in computerized form or take different forms in actual classroom situations as compared to computer-assisted learning.

The present study still throws an alarming light on objectivity in testing practices. Based on the present results we have to expect that not only the test administrator's ethnicity, but also his or her gender, may have systematic effects on self-estimations. The study indicates that men and women as test administrators and most likely also (new) teachers are systematically perceived in different ways. First, this highlights problems of objectivity for testing, as there are almost no gender-neutral situations. Second, the results point towards systematic differences in metacognitive experiences with reference to test administrators' gender. This effect has implications for learning situations as those experiences shape expectations and goals for future learning processes (Efklides 2008).

To increase knowledge, more attention has to be drawn toward the social context and characteristics of the testing and learning situations. Applying current theories from research in social psychology as well as from testing practices may lead to the formulation of hypotheses and to a better understanding of the factors that influence performance in various situations.

**Acknowledgements**

We would like to thank Eva Weißkopf and Anna Lewin for their help with the data collection and three anonymous reviewers for their thoughtful comments.

## 2.6 References

Aspinwall, L. G. (1998). Rethinking the role of positive affect in self-regulation. *Motivation and Emotion, 22*, 1-32.

Beloff, H. (1992). Mother, father and me: Our IQ. *The Psychologist, 5,* 309-311.

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41,* 174-181.

Bennett, M. (1996). Men's and women's self-estimates of intelligence. *Journal of Social Psychology, 136,* 411-412.

Bundesministerium für Familie, Senioren, Frauen und Jugend (2005). *Gender Datenreport* [Gender data report] [Electronic Version]. Retrieved January 2, 2009, from the World Wide Web http://www.bmfsfj.de/bmfsfj/generator/Publikationen/genderreport/01-Redaktion/PDF-Anlagen/gesamtdokument,property=pdf,bereich=genderreport, sprache=de,rwb=true.pdf

Boekaerts, M., & Rozendaal, J. S., (in press). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction,* doi:10.1016/j.learninstruc.2009.03.002

Bookout, D. V. T., & Hosford, R. E. (1969). Administration effects on the S-329 of the GATB using three experimental treatments. *Journal of Employment Counseling, 6,* 124-132.

Carli, L. L. (2001). Gender and social influence. *Journal of Social Issues, 57,* 725-741.

Carli, L. L., & Eagly, A. (2001). Gender, hierarchy, and leadership: An introduction. *Journal of Social Issues, 57,* 629-636.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Coleman, L. M., Jussim, L., & Abraham, J. (1987). Students' reactions to teachers' evaluations: The unique impact of negative feedback. *Journal of Applied Social Psychology, 17,* 1051-1070.

Danso, H. A., & Esses, V. M. (2001). Black experimenters and the intellectual test performance of white participants: The tables are turned. *Journal of Experimental Social Psychology, 37,* 158-165.

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

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review, 1,* 3-14.

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*, 277-287.

Efklides, A., & Petkaki, C. (2005). Effects of mood on students' metacognitive experiences. *Learning and Instruction, 15,* 415-431.

Efklides, A., Samara, A., & Petropoulou, M. (1999). Feeling of difficulty: An aspect of monitoring that influences control. *European Journal of Psychology of Education, 14,* 461-476.

Efklides, A., & Volet, S. (2005). Emotional experiences during learning: Multiple, situated and dynamic. *Learning and Instruction, 15,* 377-380.

Eurostat. (2008). *Das Leben von Frauen und Männern in Europa: Ein statistisches Porträt* [The life of men and women in Europe: A statistical portrait]. Luxembourg: European Communities.

Fein, S., & Spencer, S. J. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology, 73,* 31-44.

Fernandez-Ballesteros, R., De Bruyn, E. E. J., Godoy, A., Hornke, L. F., Ter Laak, J., Vizcarro, C., Westhoff, K., Westmeyer, H., & Zaccagnini, J.L. (2001). Guidelines for the assessment process (GAP): A proposal for discussion. *European Journal of Psychological Assessment, 17,* 187-200.

Försterling, F., Preikschas, S., & Agthe, M. (2007). Ability, luck, and looks: An evolutionary look at achievement ascriptions and the sexual attribution bias. *Journal of Personality and Social Psychology, 92,* 775-788.

Hofstede, G. (2001). Culture's consequences – Comparing values, behaviors, institutions and organizations across nations (2nd ed.). London: Thousand Oaks.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligence. *Journal of Educational Psychology, 57,* 253-270.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60,* 581-592.

Jussim, L., Coleman, L., & Nassau, S. R. (1989). Reactions to interpersonal evaluative feedback. *Journal of Applied Social Psychology, 19,* 862-884.

Katz, I., Roberts, S. O., & Robinson, J. M. (1965). Effect of difficulty, race of administrator, and instructions on Negro digit-symbol performance. *Journal of Personality and Social Psychology, 2*, 53-59.

Lynn, R., Wilberg, S., & Margraf-Stiksrud, J. (2004). Sex differences in general knowledge in German high school students. *Personality and Individual Differences, 37,* 1643-1650.

Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology, 44,* 645-657.

Marx, D. M., & Stapel, D. A. (2005). It depends on your perspective: The role of self-relevance in stereotype-based underperformance. *Journal of Experimental Social Psychology, 42,* 768-775.

Niederle, M., & Versterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quartely Journal of Economics, 122,* 1067-1101.

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin, 29,* 782-789.

Pratto, F., Stallworth, L. M., Sidanius, J., & Siers, B. (1997). The gender gap in occupational role attainment: A social dominance approach. *Journal of Personality and Social Psychology, 72,* 37-53.

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57*, 55-71.

Rammstedt, B., & Rammsayer, T. (2000). Sex differences in self-estimates of different aspects of intelligence. *Personality and Individual Differences, 29,* 869-880.

Rammstedt, B., & Rammsayer, T. (2002). Self-estimated intelligence: Gender differences, relationship to psychometric intelligence and moderating effects of level of education. *European Psychologist, 7,* 275-284.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76,* 85-97.

Reilly, J., & Mulhern, G. (1995). Gender-differences in self-estimated IQ: The need for care in interpreting group data. *Personality and Individual Differences, 18,* 189-192.

Rosenthal, R. (1976). *Experimenter effects in behavioral research.* Oxford, England: Irvington.

Rosenthal, R., & Fode, K. L. (1963). Three experiments in experimenter bias. *Psychological Reports, 12,* 491-511.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom.* New York: Rinehart and Winston.

Samuel, W. (1977). Observed IQ as a function of test atmosphere, tester expectation, and race of tester: A replication for female subjects. *Journal of Educational Psychology, 69,* 593-604.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115,* 336-356.

Schmid Mast, M. (2002). Female dominance hierarchies: Are they any different from males'? *Personality and Social Psychology Bulletin, 28,* 29-39.

Schutz, P. A., & Davis, H. A. (2000). Emotions and self-regulation during test taking. *Educational Psychologist, 35,* 243-256.

Sidanius, J., & Pratto, F. (1999). Social dominance: An intergroup theory of social hierarchy and oppression. New York: Cambridge University Press.

Sieverding, M. (2003). Women underevaluate themselves: Self-evaluation-biases in a simulated job interview. *Zeitschrift für Sozialpsychologie, 34,* 147-160.

Spencer, S. J., Fein, S., Wolfe, C. T., Fong, C., & Dunn, M. A. (1998). Automatic activation of stereotypes: The role of self-image threat. *Personality and Social Psychology Bulletin, 24,* 1139-1152.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35,* 4-28.

Steele, C. M. (1997). A threat in the air: how stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613-629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797-811.

Turner, J. C., Thorpe, P. K., & Meyer, D. K. (1997, March). Students' reports of motivation and negative affect: A theoretical and empirical analysis. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Wagner-Menghin, M. M. (2004). *Der Lexikon-Wissen-Test (LEWITE). Computergestützte Testvorgabe und Auswertung. Wiener Testsystem* [Lexical-Knowledge-Test (LKT). Computerized presentation and scoring. Vienna Testing System]. Mödling, Austria: Schuhfried.

Wheeler, S. C., Jarvis, W. B. G., & Petty, R. E. (2001). Think unto others: The self-destructive impact of negative stereotypes. *Journal of Experimental Social Psychology, 37,* 173-180.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127,* 797-826.

Willingham, W. W., & Cole, N. S. (1997). Introduction. In W. W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 1-15). Mahwah, NJ: Erlbaum.

Zanna, M. P., Sheras, P. L., Cooper, J., & Shaw, C. (1975). Pygmalion and Galatea: The interactive effect of teacher and student expectancies. *Journal of Experimental Social Psychology, 11,* 279-287.

# Chapter 3

## Too Perfect to Challenge: Effects of Attractive Examiners on Performance of Men and Women

Vormittag, I. & Ortner, T. M. (submitted). Too Perfect to Challenge: Effects of Attractive Examiners on Performance of Men and Women.

## 3.1 Abstract

We investigated effects of examiners' perceived attractiveness and examiners' gender on test performance during a standardized face-to-face testing situation assessing self-estimated and de facto verbal knowledge. One hundred fourteen nonpsychology students were individually tested by one of 22 examiners. Two results were obtained: Independent of test taker's or examiner's gender, perceived attractiveness of the administrator led to more conservative self-estimations of verbal knowledge. A moderated regression analysis further revealed a significant three-way interaction of test taker's gender, examiner's gender, and examiner's attractiveness on de facto knowledge: Men and women showed lower scores on de facto knowledge with an attractive same-gender examiner compared to their performance with an attractive opposite-gender examiner or in interaction with a nonattractive examiner.

*Keywords:* Examiner effect; Test administration effect; Perceived attractiveness; Same-gender interaction

## Too Perfect to Challenge: Effects of Attractive Examiners on Performance of Men and Women

Beginning in school and subsequently extending to working life, people typically face evaluative situations in which they are required to show knowledge or certain abilities in the presence of others. These situations range from more or less spontaneous recitations to highly standardized employment interviews or even face-to-face testing situations (e.g., Aiken & Groth-Marnat, 2006; Anastasi & Urbina, 1997; Kakkar, 2004). Such a social examination situation normally consists of at least one or several examinees and one or several examiners (Anastasi & Urbina, 1997; Domino & Domino, 2006). Demands on examined persons can therefore be described as twofold: first, to show the requested (intellectual) performance, and second, to handle the attributes of the given social interaction (e.g., Argyle, 2009; Cronbach, 1956). Achievement in such a situation will therefore depend on successful task completion as well as successfully handling the demands of the given social situation.

During an examination situation, personal values and beliefs as well as the person's perception of the situation and personal motivations have been proposed to impact the outcome as much as they do for any kind of social interaction (Hogg & Vaughan, 2008; Wittenborn, 1957). Past research has revealed that various personal attributes may influence social interactions; characteristics such as gender (Banaji & Greenwald, 1995; Burn, 1996; Eagly & Karau, 2002; Lytton & Romney, 1991), age (Gatz & Cotton, 1994; Kite, Stockdale, Whitley, & Johnson, 2005; Montepare & Zebrowitz, 1998), ethnicity (Dovidio & Gaertner, 1986; Elfenbein & Ambady, 2002; Vorauer & Kumhyr, 2001), as well as physical appearance (Anderson, John, Keltner, & Kring, 2001; Chaiken, 1979; Reis, Wheeler, & Spiegel, 1982; Swami & Furnham, 2008) have revealed relevance through eliciting attributional processes, which affect the behavior of interaction partners (Hogg & Vaughan, 2008). The current study addresses effects of examiner's gender and perceived attractiveness on test takers' achievement in a face-to-face testing situation.

### 3.2 Experimenter Effects

With reference to a possible impact of experimenters on test takers' behavior, Sattler and Theye (1967) distinguished three possible sources of such mostly undesired variance: First, so-called procedural effects address the impact of examiners' deviations from recommended standard procedures on participants' performance (e.g., Mishra, 1982). Second, situational effects have been addressed in studies investigating the influence of incentives,

rewards, or discouragement (e.g., Dickstei & Ayers, 1973; Fowler & Clingman, 1977; Sattler & Theye, 1967). Finally, in addition, the experimenter or examiner as a person has been found to be a possible source of bias. The question of examiner effects was addressed in particular in the 1960s in the field of experimental research (e.g., Graziano, Varca, & Levy, 1982; Rosenthal & Rubin, 1978; Rumenik, Capasso, & Hendrick, 1977).

Two forms of effects were further distinguished: So-called *experimenter expectancy effects* (Rosenthal, 1976) were supposed to emerge from experimenters' or examiners' attitudes and motivations. Several studies impressively revealed that experimenters' expectancies may influence persons' verbal and nonverbal behaviors in different ways and situations (Clarke, Sproston, & Thomas, 2003; Harris & Rosenthal, 1985; Judice & Neuberg, 1998; Rosenthal, 1995; Rosenthal & Jacobson, 1966).

Opposite these effects, so-called *experimenter or examiner effects* were defined as caused by perceived individual differences in the person of the experimenter, such as gender, ethnic background, or appearance (Rosenthal, 1976). The current study addresses two examiner characteristics: gender and perceived attractiveness. It therefore aims to further investigate examiner effects.

### 3.3 Effects of Experimenter's Gender on Intellectual Performance

With reference to effects on intellectual performance, past studies have addressed examiner's gender and ethnicity (e.g., Graziano et al., 1982; Huang, 2009; Katz, Roberts, & Robinson, 1965) as well as education (Yang & Yu, 2008) as possible sources of influence on a test taker's performance. Concerning experimenter's gender, Rumenik et al. (1977) concluded in their early review that male examiners elicit better performances from adult male and female test takers on achievement tasks by trend. However, results were mixed overall: For example, compared to adults, children were more clearly affected by examiner's gender and performed better overall when tested by a woman. Additionally, Rumenik et al. (1977) were concerned about limited validity as different kinds of tasks were employed in different studies. Furthermore, they criticized that only a few studies employed several male *and* female examiners, and did not control for other variables.

However, studies on the effects of examiner's gender are rare today, although applicability of early results is at least questionable, as the social meaning of gender has undergone significant changes within the last few decades (Diekman & Eagly, 2000). A recent study by Ortner and Vormittag (2011) addressed the question of gender effects when employing a standardized knowledge test in the presence of either a male or female examiner.

In advance of the knowledge task, test takers were asked to predict their performance after having seen the items. In contrast to Rumenik et al. (1977), results showed that men and women demonstrated higher performance when tested by a female administrator. No effect for de facto knowledge was found. These results are especially striking as the testing procedure was fully standardized: All examiners read the same standard instructions to the test takers, and answers on the test were recorded using a multiple choice format. However, as only a small amount of overall variance (7%) was explained, we expanded this approach. In this study we therefore followed the fully standardized approach and included another examiner characteristic as a possible additional source of variance: men's and women's attractiveness.

### 3.4 Effects of Attractiveness on Other People's Behavior

A person's perceived attractiveness has been revealed to be an important overall determinant in social interactions (cf. Dion & Stein, 1978; Swami & Furnham, 2008). Interpersonal attraction has been identified as shaped by plenty of subjective impressions, such as physical attractiveness, but also perceptions of proximity, reciprocity, and similarity (Hogg & Vaughan, 2008). Most research in the domain of attractiveness has focused on static physical attractiveness (e.g., Horai, Naccari, & Fatoullah, 1974; Reis et al., 1982), although some investigations have shown that attractiveness is a multidimensional construct and expressiveness and nonverbal behavior influence who is perceived as appealing and attracting (Friedman, Riggio, & Casella, 1988; Riggio, Widaman, Tucker, & Salinas, 1991) Under some conditions even odor contributes to overall perceived attractiveness (Foster, 2008). Perceptions of physical attractiveness could be supposed to be especially important for new acquaintances, as physical appearance is the first distinct characteristic noticed when meeting a person, especially in a standardized situation (Swami & Furnham, 2008).

Various studies found advantageous effects for attractive persons: Meta-analyses yielded that persons generally tend to ascribe more positive traits and fewer negative features to physically attractive individuals (Dion, Berscheid, & Walster, 1972; Eagly, Makhijani, Ashmore, & Longo, 1991; Langlois, Kalakanis, Rubenstein, Larson, Hallam, & Smoot, 2000). Furthermore, attractive persons are treated more positively in social interactions and have better chances in hiring situations than unattractive competitors (Dipboye, Arvey, & Terpstra, 1977; Langlois et al., 2000). Eagly et al. (1991) concluded that physical attractiveness in general provides a robust positive bias toward attractive individuals.

Despite positive attributes ascribed to attractive persons, attractiveness has also been shown to influence the course of social interactions. Studies have revealed that attractiveness – independent of expertise – is positively correlated with successful persuasion (Chaiken, 1979; Horai et al., 1974; Snyder & Rothbart, 1971; Vogel, Kutzner, Fiedler, & Freytag, 2010). Barnes and Rosenthal (1985) also found interaction effects of gender and attractiveness. For example, test takers evaluated an attractive opposite-gender examiner more positively than an attractive same-gender examiner.

Furthermore, research has indicated that perceived attractiveness can draw on attention (Maner, Gailliot, & DeWall, 2007; Maner et al., 2003; Maner et al., 2009; Sui & Liu, 2009).

There is a lack of research investigating the impact of perceived attractiveness on test performance. Karremans, Verwijmeren, Pronk, and Reitsma (2009) showed that male participants performed worse on a computerized cognitive task when they previously had contact with an attractive female examiner. No such effect was found for women in contact with a male examiner. The authors explained this by the stronger mating interests and stronger self-presentational concerns of men. The authors conclude that impression management requires cognitive resources and this led to impairment of concurrent cognitive task performance in mixed-gender interactions. Still, the question remains if perceived attractiveness influences the course of a testing procedure when examiner and examinee interact directly during assessment. Social psychological research indicates that social comparison processes with someone similar arise automatically in first impressions (Gilbert, Giesler, & Morris, 1995). Social comparisons with someone admirable – attractive and in a dominant position – could have negative consequence for the test taker (Cash, Cash, & Butters, 1983; Wood, 1989).

### 3.5 Aims of the Present Study

Based on previous studies, we investigated effects of examiners' gender and attractiveness in a standardized face-to-face testing situation.

Previous results concerning gender (Ortner & Vormittag, 2011; Rumenik et al., 1977) have been mixed. Also with reference to previous results concerning attractiveness, different effects seemed feasible: First, independent of examiners' and test takers' genders, perceived attractiveness could have an effect on test takers' performance by attracting interest and attention. An attractive examiner could therefore impair resources available to solve a task, and lead to poorer results on achievement tests.

On the other hand, attractive persons were found to be perceived as more socially competent and appealing (Langlois et al., 2000), and examiners of the opposite gender were evaluated more positively (Barnes & Rosenthal, 1985). This positive ascription could affect test takers' feelings during an examination. The mere presence of an attractive interaction partner could put the person in a positive mood. Research has shown that students' achievement motivation is heightened in a positive atmosphere (cf. Meyer & Turner, 2006; Pekrun, 1992). Following this rationale, one would expect higher scores on a test when test takers are examined by an attractive administrator.

Considering an interaction effect of gender and attractiveness, working with an attractive administrator of the opposite gender could also be distracting and result in costs of impression management (Karremans et al., 2010), whereas working with an administrator of the same gender could elicit social comparison, and if this administrator is attractive this could have negative effects on self-evaluation (Cash, Cash, & Butters, 1983; Wood, 1989). Our research design therefore allowed for the investigation of effects of gender composition, perceived attractiveness, and the interactions of these characteristics as independent variables, and performance as the dependent variable.

As in a recent study by Ortner and Vormittag (2011), we applied an adaptation of a standardized achievement test assessing general knowledge for the purpose of an oral examination. To test systematic effects of examiner's gender, two types of measures as dependent variables were included in the present study, a more *subjective* and an *objective* measure. First, a subjective score was created by the number of items a person estimated that she or he would be able to solve. Second, an objective score was employed by summing the correctly solved items supposed to be further influenced by skills and knowledge, as well as successful self-regulatory processes (Schutz and Davis 2000).

We therefore addressed the following research questions: (a) First, we investigated whether perceived attractiveness of the examiner would impact the test taker's self-estimated intellectual performance or intellectual performance independently of the administrator's and test taker's gender (main effect of attractiveness). (b) Additionally, we investigated whether there would emerge an interaction effect between gender of examiner and gender of test taker moderated by perceived attractiveness of the examiner on test performance.

## 3.6 Method

### 3.6.1 Participants

One hundred fourteen (nonpsychology) students participated as test takers (61 women and 53 men, aged 19 to 36, *M* = 24.21, *SD* = 4.02). Participation was voluntary. Test takers were blind to the research aims. Twenty-two advanced psychology students participated as examiners (11 women and 11 men, aged 21 to 36, *M* = 25.35, *SD* = 4.56). All administrators were trained by the first author regarding how to give the standardized instructions. They received an expense allowance of €8 per hour. Test takers as well as examiners were told that the purpose of this study was to investigate an originally computer-based test in a paper-pencil form.

### 3.6.2 Materials

**Self-estimation and general knowledge.** For assessment of *self-estimation of knowledge* and *de facto knowledge*, a shortened paper-pencil version of the computerized Lexical Knowledge Test (LKT; Wagner-Menghin 2004) was applied. The LKT is a Rasch-homogeneous test that consists of two parts: First, a list of words is presented and the test taker has to indicate which words he/she knows and is able to explain. This part of the test battery assesses self-estimated knowledge. Second, for each word on the list, a definition with two missing words is presented. The test taker has to choose from a list of options the missing words that complete the definition. This second part assesses de facto knowledge. The LKT second module therefore assesses crystallized intelligence (gc; Horn and Cattell 1966), the items represent different fields of verbal general knowledge, for example, art (‘‘copper engraving’’), medicine (‘‘oligophrenia’’), and nutrition (‘‘calvados’’). The 10 items used were chosen from items that fell within the medium-difficulty range. The raw scores were calculated as the total number of solved items. Homogeneity of items is given since it fulfils the criteria of the Rasch Model (Rasch, 1960).

**Attractiveness.** We assessed perceived attractiveness of the examiner with three items not focused exclusively on physical attractiveness but with regard to overall attraction (“I can imagine that the administrator is appealing to many people”; “The administrator gives a pleasant impression”; “I can imagine that it could be nice to meet the administrator privately”). Internal consistency (Cronbach’s $\alpha$) of all items was .71 in our sample.

### 3.6.3 Procedure

Test takers were recruited on the campus of the university throughout the semester. They were informed that the test assesses lexical knowledge. Participation took 15 min on average. The test takers were led individually to a quiet room of the department where they were welcomed by either a male or female examiner. The testing took place in a double-blind face-to-face setting. Examiners read written instructions prior to testing. Each administrator tested five to eight test takers subsequently. After conducting the test, the administrator asked the test taker to fill out a questionnaire concerning perceptions of the testing situation, including the items concerning perceived attractiveness of the examiner. If the test taker agreed, the completion was done in the same room but behind a movable wall. Questionnaires were not handed to administrators, but sealed in an envelope.

### 3.6.3 Statistical Analysis

To investigate effects of attractiveness we employed moderated hierarchical regression analyses[1], including attractiveness, gender of examiner, and gender of test taker, as well as all interactions between those three variables as predictors for (a) self-estimated knowledge and (b) de facto knowledge.

We entered gender of participant, gender of administrator, and attractiveness as single predictors (Block 1). In Block 2, we stepwise entered the two-way interactions, and in Block 3, the three-way interaction of attractiveness, test taker's gender, and examiner's gender. The score of attractiveness was centered in advance and the variables test taker's gender and examiner's gender were dummy-coded (0 for female and 1 for male). As a significant interaction emerged, we conducted a simple slope analysis using IRSE (Meier, 2008), reporting unstandardized regression weights.

[1] We additionally applied multilevel analyses with examiner ID as Random Intercept. Neither the Wald Statistik nor the chi-square test indicated an effect of the individual examiner. Therefore we used a hierarchical regression analysis, because a small sample size is especially prejudicial in multilevel analysis (Raudenbush & Byrk, 2002).

### 3.7 Results

Descriptive statistics of the four groups are given in Table 3.1.

*Table 3.1.* Means and Standard Deviations for Self-Estimated Knowledge and De Facto Knowledge by Group

| | Examiner | | | |
|---|---|---|---|---|
| | Woman | | Man | |
| Test taker | Woman | Man | Woman | Man |
| Self-estimated knowledge | 12.48 (2.43) | 13.19 (2.68) | 12.69 (1.80) | 12.92 (2.79) |
| De facto knowledge | 10.14 (2.55) | 12.22 (2.24) | 11.44 (2.09) | 11.85 (2.60) |

(a) Hierarchical regression analysis for self-estimated knowledge revealed attractiveness as a significant predictor ($\beta = -.24$; $p \leq .05$, $R^2 = .06$). The more attractive both women and men were perceived as examiners, the lower test takers estimated the number of items they would solve correctly. No two-way or three-way interaction reached statistical significance.

(b) Analysis including de facto knowledge as the dependent variable resulted in no significant two-way interaction, but a significant three-way interaction ($\beta = -.29$; $p \leq .05$, $\Delta R^2 = .03$; see Table 3.2). When working with examiners perceived as less attractive, test takers showed similar results with male or female examiners (see Figure 3.1). However, when attractiveness of examiners was estimated as high, women gained similar results when tested by male examiners as in the low attractiveness groups, but gained lower results with a female examiner. For male test takers the same pattern emerged: When working with attractive female examiners, male test takers performed similarly to the low attractiveness groups. When working with attractive male examiners, male test takers' performance decreased. (see Figure 3.2). Simple slope analyses revealed a significant slope for male test takers with a male examiner indicating a negative linear effect of attractiveness on de facto knowledge ($B = -2.31$, $t = -2.26$, $p = .025$). The simple slope for women tested by a female examiner showed the same trend, but was not significant. This final model explained 19% of the variance.

*Table 3.2.* Summary of Hierarchical Regression Analysis for Variables Predicting De facto Knowledge (N = 114)

| Variable | B | SE B | β |
|---|---|---|---|
| Step 1 | | | |
| Examiner's gender (TAG) | 0.34 | 0.44 | 0.07 |
| Test taker's gender (TTG) | 1.15 | 0.44 | 0.23* |
| Attractiveness (A) | -0.61 | 0.24 | -0.23* |
| Step 2 | | | |
| Examiner's gender (TAG) | 1.06 | 0.62 | 0.22 |
| Test taker's gender (TTG) | 1.95 | 0.64 | 0.40* |
| Attractiveness (A) | -0.42 | 0.37 | -0.16 |
| TAG x TTG | -1.53 | 0.89 | -0.26 |
| TAG x A | -0.58 | 0.52 | -0.12 |
| TTG x A | 0.06 | 0.48 | 0.02 |
| Step 3 | | | |
| Examiner's gender (TAG) | 1.01 | 0.61 | 0.21 |
| Test taker's gender (TTG) | 1.83 | 0.63 | 0.37* |
| Attractiveness (A) | -0.74 | 0.39 | -0.28 |
| TAG x TTG | -1.54 | 0.88 | -0.26 |
| TAG x A | 0.37 | 0.68 | 0.08 |
| TTG x A | 0.72 | 0.57 | 0.18 |
| TAG x TTG x A | -2.18 | 1.03 | -0.29* |

*Note.* $R^2 = .12$ for Step 1; $\Delta R^2 = .03$ for Step 2; $\Delta R^2 = .03$ ($p$s < .05).

*$p < .05$.

*Figure 3.1.* Interaction of examiner's gender and test taker's gender at *low* perceived attractiveness of examiner predicting de facto knowledge.

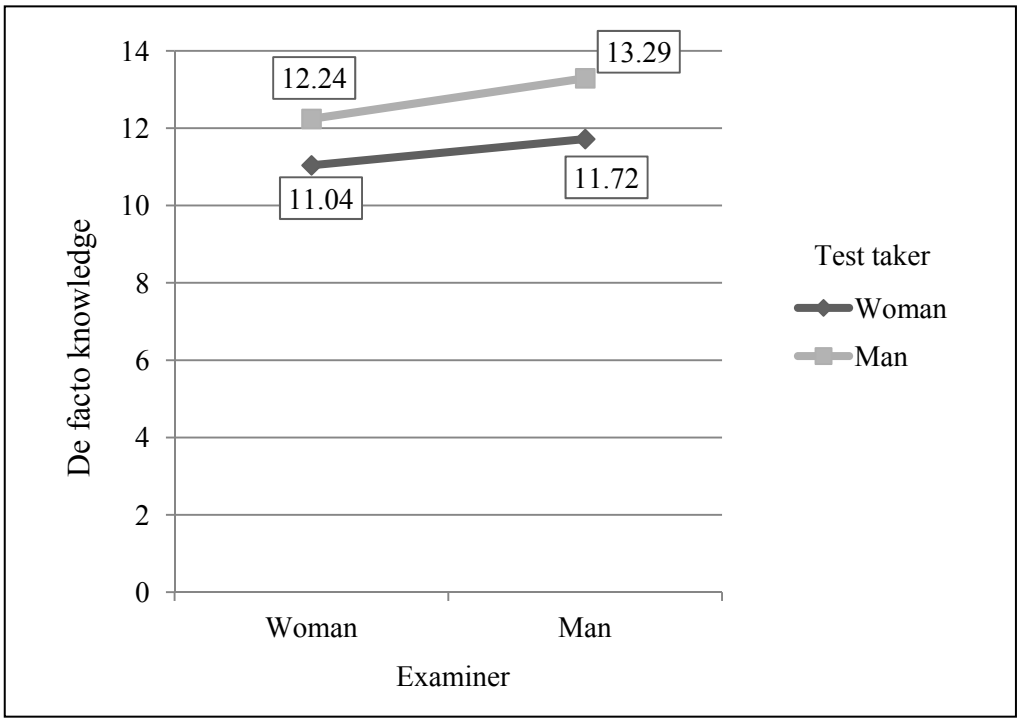
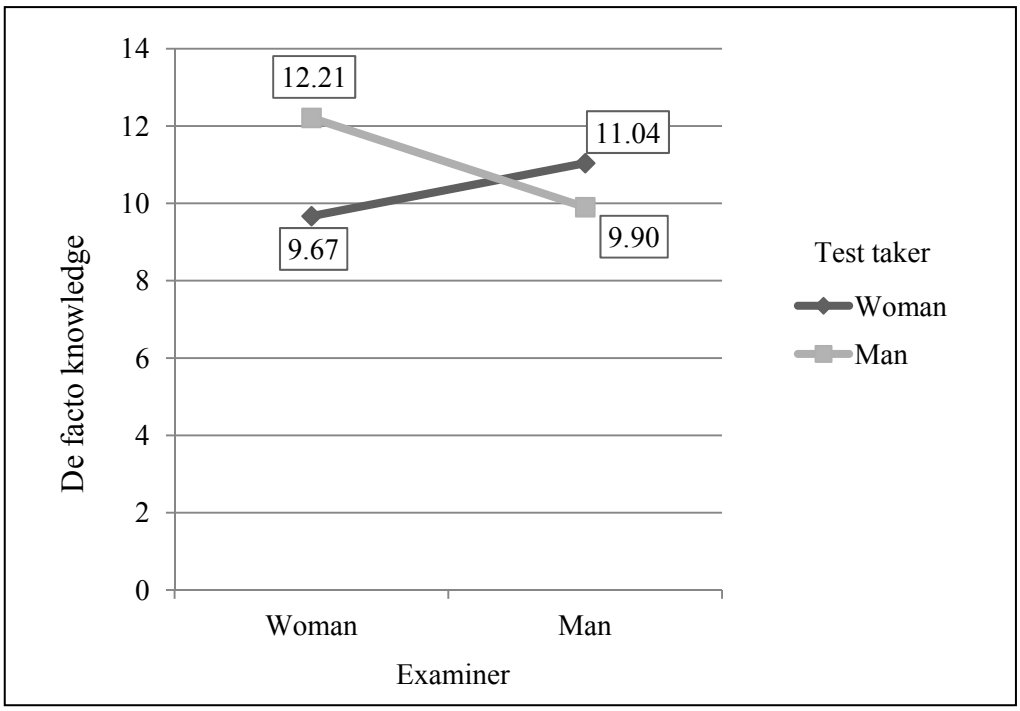



*Figure 3.2.* Interaction of examiner's gender and test taker's gender at *high* perceived attractiveness of examiner predicting de facto knowledge.



### 3.8 Discussion

The present study addressed single and combined effects of examiners' perceived attractiveness as well as examiners' gender during a standardized face-to-face testing situation. First, this study revealed a general main effect of examiners' attractiveness on participants' self-estimated knowledge. Higher perceived attractiveness of examiners was accompanied by lower predictive self-estimations of intellectual performance. In our study, confrontation with someone attractive and possibly admirable – maybe good looking and/or holding an aspiring position – led test takers to more cautious and conservative estimations of their knowledge. This result is in line with research showing negative contrast effects by lowered self-estimation in the presence of attractive others, and has been explained by an increase of self-consciousness and increased social anxiety in the presence of attractive others (e.g., Thornton & Moore, 1993). This form of *defensive pessimism* (Norem & Cantor, 1986) has also been suggested as part of a coping strategy: to underestimate performance in order to increase motivation on the test or to protect self-esteem (e.g., Elliot & Church, 2003).

With reference to de facto task performance, our study is the first to demonstrate that the interaction of three variables – test taker's gender, examiner's gender, and examiner's attractiveness – significantly predicts a person's intellectual performance, explaining 19% of the overall variance. Data revealed lower results on the knowledge task for same-gender dyads when examiners were perceived as attractive compared to test takers examined by either a nonattractive or an opposite-gender examiner.

There is a large number of existing research studies concerning upward and downward social comparison processes and the consequences of such comparisons on feelings and behavior (Blanton, Buunk, Gibbons, & Kuyper 1999; Festinger, 1954; Gilbert, Giesler, & Morris, 1995; Kruglanski & Mayseless, 1990; Suls & Wheeler, 2000; Taylor & Lobel, 1989; Tesser, Millar, & Moore, 1988; Wood, 1989). Based on social comparison theory (Festinger, 1954), two different explanations for the given results seem feasible. Several studies have suggested that exposure to attractive images may affect men's and women's self-evaluations. Thornton and Moore (1993) showed that comparisons with either attractive or unattractive same-gender targets led to contrast effects in men's and women's self-evaluations: Upward comparison with attractive targets evoked negative evaluations of one's own attractiveness, whereas comparison with unattractive targets led to more positive self-evaluations. These effects were replicated for women and men after confrontation with highly attractive same-gender images (see Groesz, Levine, & Murnen, 2002; Hargreaves & Tiggemann, 2009). If compared with attractive and seemingly successful same-gender students, such negative

contrast effects could have occurred in our study as well. One possible explanation for our results is that negative self-evaluation and a preoccupation with negative social comparison have detached cognitive capacity: Those participants engaged in contrastive comparisons were not able to fully focus on the lexical knowledge task and showed lowered results.

It is also possible that the devaluing social comparison elicited behavioral contrast effects. Only a few studies have addressed the effects of social comparison on actual behavior (Pelham & Wachsmuth, 1995; Stapel & Suls, 2004). Recently, Stapel and Koomen (2000, 2001) introduced the *interpretation comparison model*. They assume social comparison to elicit two different processes: On the one hand, information can be used in interpretational terms for defining the self. This interpretational process leads to assimilation. On the other hand, the information can be used as a comparative standard against which the self is evaluated. This comparative process leads to contrast effects. Stapel and Suls (2004) found support for this model in a series of studies: Participants who were explicitly asked to compare themselves with another person searched for similarities and compared themselves on a specific dimension, which instigated assimilation. When participants implicitly compared themselves with a target, contrast effects arose. In the context of upward social comparison, this suggests that explicit comparison may lead to more positive self-evaluation and better performance, whereas implicit comparison may evoke negative self-evaluation and performance decrements. So far, studies have applied the implicit comparison only in priming procedures with extreme comparison targets (cf. Stapel & Suls, 2004). It remains unclear whether mere interaction with an admirable comparison target can elicit such a behavioral contrast as well. However, Gilbert, Giesler, and Morris (1995) claimed that social comparisons in real life happen spontaneously. Blanton and Stapel (2008) further corroborated that contrastive effects arise in situations in which individuals compare their personal selves with a target. In line with these assumptions, we may conclude that in our testing situation, processes of implicit social comparison with a similar target could be instigated.

Why did the contrast effect occur only for same-gender interactions? Spontaneous social comparisons that are often outside of awareness do not occur in all interactions. One precondition for the emergence of social comparisons is similarity or comparability between oneself and the target; gender has been described as one of the major dimensions that indicate similarity (Tesser, 1986; Tesser & Campbell, 1983). So it seems probable that the upward comparison with an attractive, same-gender target in an aspiring position, namely an

examiner, led to contrast effects: Participants performed worse than in the other conditions, where either no relevant comparison target was present, or the target was not admirable.

Another explanation for the differential effects of examiners in relation to their perceived attractiveness may be differences in their behavior. Rosenthal (1976) revealed nonverbal cues that influence the test taker, and even standardized instruction and administration cannot rule out that examiners behaved differently. Furthermore, research has shown that perceived attractiveness is related to expressiveness (cf. Riggio et al., 1991).

In contrast to the previous finding by Ortner and Vormittag (2011), we did not reveal a main effect of gender. This is surprising, as we employed a similar setting as well as the same testing materials. However, the main difference between the studies lies in the age difference between examiners and test takers. Whereas in the cited study examiners were graduate students in psychology at the end of their study, the current study mainly employed bachelor students at the end of courses in Psychological Assessment. We therefore propose a possible age or additional status interaction effect of examiners such that older examiners will elicit effects different from examiners who are perceived as similar in age; this should be addressed in future studies.

There are limitations of the results: Participation was voluntary; therefore, possible negative consequences of poor test taking performance were not as evident as they would be in a real examination situation. However, we would still expect strong possibilities of ego threat in a face-to-face testing situation such as we investigated. Furthermore, this study included only one particular facet of intelligence, namely, verbal knowledge. Future studies can therefore increase data with reference to situations and test materials. Furthermore, one could argue that asking persons about attractiveness after a testing situation may be influenced by more variables than solely examiners' perceived attractiveness, such as, for example, by derogative effects of highly attractive same-gender individuals (e.g., Maner et al., 2007; Maner, Miller, Rouby, & Gailliot, 2009; Sui & Liu, 2009), or motivated stereotyping after a potentially ego-threatening testing situation (Sinclair & Kunda, 2000). However, as we already found an effect applying this design, we may expect even larger effects in the case of assessing examiners' attractiveness more implicitly. Furthermore, additional effects may lead to certain intellectual decrements, especially in an examination situation, and explain further variance, such as, for example, *stereotype threat* (cf. Steele, 1997), or even stereotype priming (e.g., Ortner & Sieverding, 2008) through the presence of an examiner. Further studies should also address these questions.

With reference to practical implications, our results show that today, even when using standardized testing procedures, individual characteristics of test administrators or examiners may systematically influence test takers' self-estimations and performance. We employed a multiple-choice standardized test with fully standardized instructions. Due to administrators' attractiveness and gender, verbal knowledge of students was malleable. In contrast to early studies (see Rumenik et al., 1977), 22 persons worked as examiners. Influences of individual characteristics of the examiners seem therefore negligible.

One basic claim of standardized psychological assessment instruments is objectivity in terms of guaranteeing results that are independent of the examiner's characteristics and behavior (Stuart-Hamilton, 1996; Westmeyer, 2003). The present study proposes that in the case of oral examination, standardization cannot guarantee objective assessment. As administrators' characteristics affect test takers' performance in such a standardized situation, this basic claim would be violated (see Aiken & Groth-Marnat, 2006). In fact, oral examinations had been criticized earlier for low reliability and validity compared to written examinations (Daelmans, Scherpbier, Van der Fleuten, & Donker, 2001; Pokorny & Frazier, 1966). However, there is not a clear solution yet as previous studies have revealed administrator effects even in computerized settings (see Karremans et al., 2009). Still, our results question oral exams in general, although certain advantages of oral examinations cannot be replaced by written examinations: The interactive social situation provides additional information such as appearance or presentational style; faking and cheating are more difficult than on written examinations (Aiken & Groth-Marnat, 2006).

In summary, the present study enriches the existing research on factors influencing performance estimations and task performance in face-to-face testing situations with reference to administrators' attractiveness and gender. To further increase knowledge of social effects on self-estimation and performance, more attention has to be drawn toward the social context and characteristics of the testing situation including settings in real examination practice.

### 3.9 References

Aiken, L. R., & Groth-Marnat, G. (2006). *Psychological Testing and Assessment* (12th ed.). Boston: Pearson Allyn & Bacon.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River (NJ): Prentice Hall.

Anderson, C., John, O. P., Keltner, D., & Kring, A. M. (2001). Who attains social status? Effects of personality and physical attractiveness in social groups. *Journal of Personality and Social Psychology, 81*, 116-132. doi: 10.1037/0022-3514.81.1.116

Argyle, M. (2009). *Social interaction* (2nd ed.). Oxford England: Atherton Press.

Banaji, M. R., & Greenwald, A. G. (1995). Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology, 68*, 181-198. doi: 10.1037/0022-3514.68.2.181

Barnes, M. L., & Rosenthal, R. (1985). Interpersonal effects of experimenter attractiveness, attire, and gender. *Journal of Personality and Social Psychology, 48*, 435-446. doi: 10.1037/0022-3514.48.2.435

Blanton, H., Buunk, B. P., Gibbons, F. X., & Kuyper, H. (1999). When better-than-others compares upward: the independent effects of comparison choice and comparative evaluation on academic performance. *Journal of Personality and Social Psychology, 76*, 420–430. doi: 10.1037/0022-3514.76.3.420

Blanton, H., & Stapel, D. A. (2008). Unconscious and spontaneous and … complex: the three selves model of social comparison assimilation and contrast. *Journal of Personality and Social Psychology, 94*, 1018-1032. doi: 10.1037/0022-3514.94.6.1018

Burn, S. M. (1996). *The Social Psychology of Gender*. New York (NY): McGraw-Hill.

Cash, T. F, Cash, D. W., & Butters, J. W. (1983). "Mirror, Mirror, on the Wall…?": Contrast Effects and Self-Evaluations of Physical Attractiveness. *Personality and Social Psychology Bulletin*, 9, 351-358. doi: 10.1177/0146167283093004

Chaiken, S. (1979). Communicator physical attractiveness and persuasion. *Journal of Personality and Social Psychology, 37*, 1387-1397. doi: 10.1037/0022-3514.37.8.1387

Clarke, P., Sproston, K., & Thomas, R. (2003). An investigation into expectation-led interview effects in health surveys. *Social Science and Medicine, 56*, 2221-2228. doi: 10.1016/S0277-9536(02)00238-1

Cronbach, L. J. (1956). Assessment of individual differences. *Annual Review of Psychology, 7*, 173-196. doi: 10.1146/annurev.ps.07.020156.001133

Daelmans, H. E. M., Schierpbier, A. J. J. A., Van der Vleuten, C. P. M., & Donker, A. J. M. (2001). Reliability of clinical oral examinations re-examined. *Medical Teacher, 23*, 422-424. doi: 10.1080/01421590120042973

Dickstei, L. S., & Ayers, J. (1973). Effect of an incentive upon intelligence test performance. *Psychological Reports, 33*, 127-130.

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin*, *26*, 1171-1188. doi: 10.1177/0146167200262001

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. Journal of Personality and Social Psychology, 24, 285–290. doi: 10.1037/h0033731

Dion, K. K., & Stein, S. (1978). Physical attractiveness and interpersonal influence. *Journal of Experimental Social Psychology, 14*, 97-108. doi: 10.1016/0022-1031(78)90063-X

Dipboye, R. L., Arvey, R. D., & Terpstra, D. E. (1977). Sex and physical attractiveness of raters and applicants as determinants of resume evaluations. *Journal of Applied Psychology, 62*, 288-294. doi: 10.1037/0021-9010.62.3.288

Domino, G., & Domino, M. L. (2006). *Psychological Testing: An Introduction* (2nd ed.): Cambrigde University Press.

Dovidio, J. F., & Gaertner, S. L. (1986). *Prejudice, discrimination, and racism*. London: Academic Press, Inc.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*, 573-598. doi: 10.1037/0033-295X.109.3.573

Eagly, A. H., Makhijani, M. G., Ashmore, R. D., & Longo, L. C. (1991). What is beautiful is good, but - a meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*, 109-128. doi: 10.1037/0033-2909.110.1.109

Elfenbein, H. A., & Ambady, N. (2002). Is there an in-group advantage in emotion recognition? *Psychological Bulletin, 128*, 243-249. doi: 10.1037/0033-2909.128.2.243

Elliot, A.J. & Church, M. A. (2003). A motivational analysis of defensive pessimism and self-handicapping. *Journal of Personality*, *71*, 233-249. doi: 10.1111/1467-6494.7103005

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117–140. doi: 10.1177/001872675400700202

Foster, J. D. (2008). Beauty is Mostly in the Eye of the Beholder: Olfactory Versus Visual Cues of Attractiveness. *Journal of Social Psychology*, 148, 765-773.

Fowler, R. L., & Clingman, J. (1977). The influence of intrinsic and extrinsic reward on intratest performance of high-scoring and low-scoring children. *Psychological Record, 27*, 603-610.

Friedman, H. S., Riggio, R. E., & Casella, D. F. (1988). Nonverbal skill, personal charisma, and initial attraction. *Personality and Social Psychology Bulletin*, 14, 203-211. doi: 10.1177/0146167288141020

Gatz, M., & Cotton, B. (1994). Age as a dimension of diversity: the experience of being old. In E. J. Trickett, R. J. Watts & D. Birman (Eds.), *Human Diversity: Perspectives on people in context*. San Francisco, CA: Jossey-Bass.

Gilbert, D. T., Giesler, R. B., & Morris, K. A. (1995). When comparisons arise. *Journal of Personality and Social Psychology, 69*, 227–236. doi: 10.1037/0022-3514.69.2.227

Graziano, W. G., Varca, P. E., & Levy, J. C. (1982). Race of examiner effects and the validity of intelligence-tests. *Review of Educational Research, 52*, 469-497. doi: 10.2307/1170263

Groesz, L. M., Levine, M. P., & Murnen, S. K. (2002). The effect of experimental presentation of thin media images on body satisfaction: A meta-analytic review. *International Journal of Eating Disorders, 31,*1–16. doi: 10.1002/eat.10005

Hargreaves, D. A., & Tiggemann, M. (2009). Muscular Ideal Media Images and Men's Body Image: Social Comparison Processing and Individual Vulnerability. *Psychology of Men & Masculinity*, 10, 109-119. doi: 10.1037/a0014691

Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects - 31 meta-analyses. *Psychological Bulletin, 97*, 363-386. doi: 10.1037/0033-2909.97.3.363

Hogg, M. A., & Vaughan, G., M. (2008). *Social Psychology* (5th ed.). Harlow: Pearson Education.

Horai, J., Naccari, N., & Fatoullah, E. (1974). The effect of expertise and physical attractiveness upon opinion agreement and liking. *Sociometry, 37*, 601-606. doi: 10.2307/2786431

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligence. *Journal of Educational Psychology, 57*, 253 - 270. doi: 10.1037/h0023816

Huang, M. H. (2009). Race of the interviewer and the black-white test score gap. *Social Science Research, 38*, 31-40. doi: 10.1016/j.ssresearch.2008.07.004

Judice, T. N., & Neuberg, S. L. (1998). When interviewers desire to confirm negative expectations: Self-fulfilling prophecies and inflated applicant self-perceptions. *Basic and Applied Social Psychology, 20*, 175-190. doi: 10.1207/s15324834basp2003_1

Kakkar, S. B. (2004). *Educational Psychology*. New Dehli: Prentice-Hall of India Pvt.Ltd.

Karremans, J. C., Verwijmeren, T., Pronk, T. M., & Reitsma, M. (2009). Interacting with women can impair men's cognitive functioning. *Journal of Experimental Social Psychology, 45*, 1041-1044. doi: 10.1016/j.jesp.2009.05.004

Katz, I., Roberts, S. O., & Robinson, J. M. (1965). Effects of task-difficulty, race of administrator, and instructions on digit-symbol performance of Negroes. *Journal of Personality and Social Psychology, 2*, 53-59. doi: 10.1037/h0022080

Kite, M. E., Stockdale, G. D., Whitley, B. E., & Johnson, B. T. (2005). Attitudes toward younger and older adults: An updated meta-analytic review. *Journal of Social Issues, 61*, 241-266. doi: 10.1111/j.1540-4560.2005.00404.x

Kruglanski, A.W., & Mayseless, O. (1990). Classic and current social comparison research: expanding the perspective. *Psychological Bulletin*, *108*, 195-208. doi: 10.1037/0033-2909.108.2.195

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*, 390-423. doi: 10.1037/0033-2909.126.3.390

Lytton, H., & Romney, D. M. (1991). Parents Differential Socialization of Boys and Girls - a Metaanalysis. *Psychological Bulletin, 109*, 267-296. doi: 10.1037/0033-2909.109.2.267

Maner, J. K., Gailliot, M. T., Rouby, D. A., & Miller, S. L. (2007). Can't take my eyes off you: Attentional adhesion to mates and rivals. *Journal of Personality and Social Psychology, 93*, 389-401. doi: 10.1037/0022-3514.93.3.389

Maner, J. K., Kenrick, D. T., Becker, D. V., Delton, A. W., Hofer, B., Wilbur, C. J., & Neuberg, S. L. (2003). Sexually selective cognition: Beauty captures the mind of the beholder. *Journal of Personality and Social Psychology*, 85, 1107–1120. doi: 10.1037/0022-3514.85.6.1107

Maner, J. K., Miller, S. L., Rouby, D. A., & Gailliot, M. T. (2009). Intrasexual vigilance: the implicit cognition of romantic rivalry. *Journal of Personality and Social Psychology, 97*, 74-87. doi: 10.1037/a0014055

Meier, L. L. (2008). IRSE. Interactions in Multiple Linear Regression with SPSS and Excel (Version 1.6) [Computer software and manual]: Retrieved 23.7.2010 from http://www.urenz.ch/irse.

Meyer, D. K., & Turner, J. C. (2006). Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review, 18*, 377-390. doi: 10.1007/s10648-006-9032-1

Mishra, S. P. (1982). Intelligence-test performance as affected by anxiety and test administration procedures. *Journal of Clinical Psychology, 38*, 825-829. doi: 10.1002/1097-4679(198210)38:4<825::AID-JCLP2270380423>3.0.CO;2-2

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: the salience and significance of age in social judgments. *Advances in Experimental Social Psychology, 30*, 93-161.

Norem, J. K. & Cantor, N. (1986). Anticipatory and post hoc cushioning strategies: optimism and defensive pessimism in "risky" situations. *Cognitive Therapy and Research*, *10,* 347-362. doi: 10.1007/BF01173471

Ortner, T. M., & Sieverding, M. (2008). Where are the gender differences? Male priming boosts spatial skills in women. *Sex Roles, 59*, 274-281. doi: 10.1007/s11199-008-9448-9

Ortner, T. M., & Vormittag, I. (2011). Test administrator's gender affects female and male students' self-estimated verbal general knowledge. *Learning and Instruction, 21,* 14-21. doi: 10.1016/j.learninstruc.2009.09.003

Pekrun, R. (1992). The Impact of Emotions on Learning and Achievement - Towards a Theory of Cognitive Motivational Mediators. *Applied Psychology-an International Review-Psychologie Appliquee-Revue Internationale, 41*, 359-376. doi : 10.1111/j.1464-0597.1992.tb00712.x

Pelham, B. W., & Wachsmuth, J. O. (1995). The waxing and waning of the social self: Assimilation and contrast in social comparison. *Journal of Personality and Social Psychology, 69,* 825–838. doi: 10.1037/0022-3514.69.5.825

Pokorny, A. D., & Frazier, S. H. (1966). An Evaluation of Oral Examinations. *Journal of Medical Education, 41*, 28-40.

Raudenbush, S. W., & Byrk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Pedagogiske Institut.

Reis, H. T., Wheeler, L., Spiegel, N, Kernis, M. H., Nezlek, J., & Perri, M. (1982). Physical attractiveness in social interaction: II. Why does appearance affect social experience? *Journal of Personality and Social Psychology 43,* 979-996. doi: 10.1037/0022-3514.43.5.979

Riggio, R. E., Widaman, K. F., Tucker, J. S., & Salinas, C. (1991). Beauty is More Than Skin Deep: Components of Attractiveness. *Basic and Applied Social Psychology, 12,* 423-439. doi: 10.1207/s15324834basp1204_4

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research*. New York, N.Y.: Irvington Publishers, Inc.

Rosenthal, R. (1995). Critiquing Pygmalion: a 25-Year Perspective. *Current Directions in Psychological Science, 4*, 171-172. doi: 10.1111/1467-8721.ep10772607

Rosenthal, R., & Jacobson, L. (1966). Teachers Expectancies – Determinants of Pupils' IQ Gains. *Psychological Reports, 19,* 115

Rosenthal, R. , & Rubin, D. B. (1978). Interpersonal Expectancy Effects – 1st 345 studies. *Behavioral and Brain Sciences, 1*, 377-386. doi: 10.1017/S0140525X00075506

Rumenik, D. K., Capasso, D. R., & Hendrick, C. (1977). Experimenter Sex Effects in Behavioral Research. *Psychological Bulletin, 84,* 852-877. doi: 10.1037/0033-2909.84.5.852

Sattler, J. M., & Theye, F. (1967). Procedural, Situational, and Interpersonal Variables in Individual Intelligence Testing. *Psychological Bulletin, 68*, 347-360. doi: 10.1037/h0025153

Schutz, P. A., & Davis, H. A. (2000). Emotions and self-regulation during test taking. *Educational Psychologist, 35*, 243–256. doi: 10.1207/S15326985EP3504_03

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me, *Personality and Social Psychology Bulletin, 26*, 1329-1342. doi: 10.1177/0146167200263002

Snyder, M., & Rothbart, M. (1971). Communicator attractiveness and opinion change. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 3,* 377-387. doi : 10.1037/h0082280

Stapel, D. A., & Koomen, W. (2000). Distinctness of others and malleability of selves: Their impact on social comparison effects. *Journal of Personality and Social Psychology, 79,* 1068–1087. doi: 10.1037/0022-3514.79.6.1068

Stapel, D. A., & Koomen, W. (2001). I, we, and the effects of others on me: How self-construal moderates social comparison effects. *Journal of Personality and Social Psychology, 80,* 766–781. doi: 10.1037/0022-3514.80.5.766

Stapel, D. A., & Suls, J. (2004). Method matters: Effects of implicit versus explicit social comparisons on activation, behavior, and self-views. *Journal of Personality and Social Psychology, 87*, 860–875. doi: 10.1037/0022-3514.87.6.860

Steele, C. M. (1997). A threat in the air - How stereotypes shape intellectual identity and performance. American Psychologist, 52, 613-629. doi: 10.1037/0003-066X.52.6.613

Stuart-Hamilton, I. (1996). Dictionary of psychological testing, assessment, and treatment (Rev. ed.). London:Kingsley

Sui, J.& Liu, C. H. (2009).Can beauty be ignored? Effects of facial attractiveness on covert attention. *Psychonomic Bulletin & Review, 16*, 276-281. doi: 10.3758/PBR.16.2.276

Suls, J., & Wheeler, L. (Eds.). (2000). *Handbook of social comparison: Theory and research*. Dordrecht, NL: Kluwer.

Swami, V., & Furnham, A. (2008). The psychology of physical attraction. London: Routledge.

Taylor, S. E., & Lobel, M. (1989). Social comparison activity under threat: downward evaluation and upward contacts. *Psychological Review, 96*, 569–575. doi: 10.1037/0033-295X.96.4.569

Tesser, A. (1986). Some effects of self-evaluation maintenance on cognition and action. In R. M. Sorrentino & E. T. Higgins (Eds.), *The handbook of motivation and cognition: Foundations of social behav*ior( pp. 435-464). New York: Guilford Press.

Tesser, A., & Campbell, J. (1983). Self-definition and self-evaluation maintenance. In J. Suls & A. Greenwald (Eds.), *Social psychological perspectives on the self (Vol.* 2), pp. 1-31).

Tesser, A., Millar, M., & Moore, J. (1988). Some affective consequences of social comparison and reflection processes: The pain and pleasure of being close. *Journal of Personality and Social Psychology, 54,* 49–61. doi: 10.1037/0022-3514.54.1.49

Thornton, B., & Moore, S. (1993). Physical Attractiveness Contrast Effect: Implications for Self-Esteem and Evaluations of the Social Self. *Personality and Social Psychology Bulletin, 19,* 474–480. doi: 10.1023/A:1018867409265

Vogel, T., Kutzner, F., Fiedler, K., & Freytag, P. (2010). Exploiting Attractiveness in Persuasion: Senders' Implicit Theories About Receivers' Processing Motivation.

*Personality and Social Psychology Bulletin, 36*(6), 830-842. doi: 10.1177/0146167210371623

Vorauer, J. D., & Kumhyr, S. M. (2001). Is this about You or Me? Self-Versus-Other-Directed Judgments and Feelings in Response to Intergroup Interaction. *Personality and Social Psychology Bulletin, 27,*706–719. doi: 10.1177/0146167201276006

Wagner-Menghin, M. M. (2004). Der Lexikon-Wissen-Test (LEWITE). Computergestützte Testvorgabe und Auswertung. Wiener Testsystem) [Lexical-Knowledge-Test (LKT). Computerized presentation and scoring. Vienna Testing System]. Mödling: Schuhfried.

Westmeyer, H. (2003). Objectivity. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment*. London: Sage.

Wittenborn, J. R. (1957). The theory and technique of assessment. *Annual Review of Psychology, 8*, 331-356. doi: 10.1146/annurev.ps.08.020157.001555

Wood, J. V. (1989). Theory and research concerning social comparison of personal attributes. *Psychological Bulletin, 106,* 231–248. doi: 10.1037/0033-2909.106.2.231

Yang, M. L., & Yu, R. R. (2008). The interviewer effect when there is an education gap with the respondent: Evidence from a survey on biotechnology in Taiwan. *Social Science Research, 37*, 1321-1331. doi: 10.1016/j.ssresearch.2008.05.008

# Chapter 4

## Does Gender Speak Louder than Words? How Stereotypes Influence Perceptions of and Preferences for Test Examiners

Vormittag, I. & Ortner, T. M. (submitted). Does Gender Speak Louder than Words? How Stereotypes Influence Perceptions of and Preferences for Test Examiners.

## 4.1 Abstract

We addressed potential test takers' preferences for women or men as examiners as well as how examiners were perceived depending on their gender. We collected data from 129 university students in a pilot study and then employed an online design with 375 students who provided preferences for and ratings of examiners based on short video clips. The clips showed four out of 15 psychologists who differed in age (young vs. middle-aged) and gender giving an introduction to a fictional intelligence test session. We found a significant preference for choosing women as examiners in both studies. Employing repeated measures ANOVAs, women examiners were generally rated as more socially competent, whereas no gender differences were revealed for expertise ratings. Multinomial logistic regression indicated that, in general, preferences for examiners were made based on both estimated social competence and expertise. Loglinear analysis revealed that test taker's gender did not influence preference for examiner's gender, but social dominance orientation resulted in stronger preferences for women as examiners. Results are discussed with reference to test performance and fairness.

*Keywords:* Examiner; Test administration; Test taker preferences; Perceived competence

## Does Gender Speak Louder than Words? How Stereotypes Influence Perceptions of and Preferences for Test Examiners

Today, psychological assessment, especially test use, plays a major role in educational and vocational selection and placement (e.g., Fernández-Ballesteros, 1999; Muñiz & Bartram, 2007). In the domain of achievement testing, the employment of psychological assessment procedures generally aims for the establishment of reliable, valid, and fair measurement and should provide the opportunity for test takers to show their maximum performance. Different actions have been taken to establish such an assessment process with special attention paid to standardization, ethical responsibility, and evaluation of assessment methods (cf. Guidelines for the Assessment Process; Fernández-Ballesteros et al., 2001; DIN 33430, Westhoff et al., 2005). However, in the assessment process, there may be biasing elements that have been thus far disregarded or even elements that cannot be standardized, even if substantial effort is made. One such element in the assessment process may be the person who administers the test – the examiner.

Earlier studies in the framework of experimenter effects have shown significant effects of the experimenter's characteristics or behaviour on the participant's behaviour (e.g., Harris & Rosenthal, 1985; Rosenthal, 1976). In recent years, several studies addressed examiner effects on intellectual performance on standardized achievement tests: Characteristics such as ethnicity or the attractiveness of the examiner (see Huang, 2009; Karremans, Verwijmeren, Pronk, & Reitsma, 2009; Mishra, 1980) have revealed systematic effects on tested persons' intellectual performances. With regard to gender, Ortner and Vormittag (2011) recently reported effects due to examiner's gender in a standardized test procedure: For a face-to-face knowledge task, test takers had to estimate their own results in advance. Men and women tested by a female examiner made significantly higher estimations of their results than individuals tested by a male examiner.

Although such examiner effects have been found repeatedly – indicating that examiner characteristics may influence test takers even when using standardized assessment procedures – the underlying mechanisms causing these effects have hardly been investigated. For example, there is a lack of research addressing *how* examiners are perceived by potential examinees. Investigations of perceptions of examiners' characteristics and evaluations of the examiners may help explain effects that underlie previously reported differences in the behaviours and intellectual performances of the test takers. Another unexplored topic concerns requests and preferences: Whereas in academic settings students are sometimes able

to choose between different examiners for an oral examination, in assessment procedures test takers typically cannot choose the examining person. With regard to examiner effects, the question arises as to whether test takers would prefer certain examiners, indicating systematically different expectations. The current study therefore addressed test takers' preferences and how male and female examiners are perceived with reference to two potential test-situation-related characteristics.

## 4.2 How Examiners are Perceived

From the test taker's subjective view, examiners could be seen as strangers in a powerful position: Examinees may remember from oral examination settings at school that examiners might decide or at least have an influence on an assessment's results and its consequences. Early research on social judgments and first impressions has revealed that humans use social categories in interactions to simplify the perception processes (Allport, 1954). Stereotypes, such as beliefs about characteristics of members of distinctive social groups and their belonging to certain social categories or roles, may serve as such social categories (Fiske & Taylor, 1996; Hilton & von Hippel, 1996). In fact, every person can be assigned to several social categories; however, categories that based on physical cues (e.g., age, gender, and ethnicity) often prevail in first impressions (Fiske & Taylor, 1996). Social psychological models postulate that first impressions heavily rely on stereotypical information (Fiske & Neuberg, 1990; Kunda & Thagard, 1996), and people tend to activate stereotypes, especially in self-image-threatening situations (Spencer et al., 1998), which the assessment setting may be perceived as.

## 4.3 Influences of Gender, Age, and Other Stereotypes

Although the occupational situation of many women has changed in recent decades, with an increasing number of women in more powerful positions (Diekman & Eagly, 2000; European Commission, 2011), gender stereotypes remain stable, describing women as more expressive and men as more instrumental (Spence & Buckner, 2000).

Another physical cue for stereotypical judgments is perceived age (e.g., Kite, Stockdale, Whitley, & Johnson, 2005). Despite general negative attitudes and ageism, older employees are perceived as reliable, conscientious, and effective (Posthuma & Campion, 2009; Redman & Snape, 2002).

Furthermore, social judgments are swayed by physical or vocal cues associated with certain traits. For example, sexually dressed women have been given lower ratings on expertise (e.g., Glick et al., 2005), whereas eye glasses have been associated with higher expertise, and men's beards have led to lower expertise ratings. Ko, Judd, and Stapel (2009) revealed that persons with more masculine voices – independent of actual gender – were rated as more competent.

Besides these general effects, individuals differ in their susceptibility to and reliance on stereotypical information. For example, *social dominance orientation* (SDO) has been identified as being related to approval of stereotypic views (Jost, Banaji, & Nosek, 2004; Pratto, Sidanius, Stallworth, & Malle, 1994; Sidanius & Pratto, 2001). SDO refers to the support of intergroup hierarchies and inequality between social groups. The potential relevance of SDO for individual behaviour within assessment procedures was shown by Danso and Esses (2001). In their study, White test takers with high SDO gained better results on an ability test when tested by a Black examiner compared to White test takers with a low SDO or those tested by a White examiner. Results were not explained by negative attitudes toward Blacks, but by activated intergroup competition, boosting the performance of those who strongly identified with social hierarchies.

Another relevant person characteristic influencing stereotyping behaviour is the gender of the perceiver and the interaction of perceiver's and target's gender. Basow (1995, 2000) reported that ratings of male college professors were independent of students' gender, whereas female professors were rated more positively by women and got their lowest ratings from men.

## 4.4 Aims of the Present Research and Research Questions

There is presently no literature dealing with preferences of examinees with respect to examiner characteristics. In this study, we addressed potential test takers' preferences for women or men as examiners as well as how examiners were perceived depending on their gender. We had the following aims and hypotheses:

1. We aimed to investigate whether test takers would have a preference for either male or female examiners. We hypothesized that examinees would prefer women as examiners because expected higher social competence would lead to the expectation of a more convenient test situation. Moreover, women's lower estimated proficiency was expected to facilitate downward comparisons with self-enhancing effects and the option to restore a positive self-image in a test situation of potential threat (Taylor & Lobel, 1989; Wills, 1981).

2. Based on the literature (e.g., Eckes, 2002), we aimed to determine whether female examiners were perceived as more socially competent, whereas male examiners were perceived as possessing more expertise.

3. Due to a lack of research in this area, our third aim was to increase knowledge on preferred examiner characteristics in a test situation in general, and to investigate how perceived social competence as well as expected expertise would influence whether an examiner was chosen.

4. We also aimed to investigate the impact of test taker characteristics – namely test taker gender and test taker SDO – on the preference for either a male or female examiner. Based on the literature, we expected a person with a higher SDO to have a stronger gender stereotype effect, resulting in a preference for a female examiner. Additionally, we wanted to explore the effect of test-taker gender on the evaluation without proposing an a priori hypothesis.

We first conducted a pilot study (Study 1) to investigate the preference for female or male examiners in a real-life university setting. Second, in the main study (Study 2), we employed an online study design and presented video clips of different examiners giving standardized assessment instructions.

## 4.5 Study 1

In the pilot study, we asked students after courses for help in finding persons with certain characteristics. They were told that a small number of persons were missing in the representative data collection of an ongoing project regarding aptitude testing. Besides other information, we asked the students to give their preference for a female or male examiner at the end. The choice was either (a) between male examiner, female examiner, or a third "I do not care" option or (b) in a dichotomous format including only the choices of male or female.

### 4.5.1 Materials and Method

**Participants.** One hundred twenty-nine psychology students (Condition 1: $n$ = 63; Condition 2: $n$ = 66) were asked for help in finding persons with certain characteristics to finish data collection in a running project on intelligence. Participation was voluntary and anonymous, and persons were asked to identify themselves later with a code word.

**Procedure.** After the lecture, students were asked for several characteristics in a questionnaire (including gender, age, number and ages of siblings). We told them that certain persons fulfilling the missing characteristics would be contacted by one out of two fictional examiners: Mr. Ertl or Mrs.Weber. Students were told that both had completed their study in our department some years ago and are really nice and competent. After filling out the form, test takers were informed that no testing would take place.

**Statistical analysis.** We conducted chi-square tests to test for preferences for examiner's gender and to check for differential preferences depending on participants' gender.

### 4.5.2 Results

In Condition 1, 54 participants chose the option "I do not care," nine students chose the female examiner, and no student chose the male examiner. This difference reached significance, $\chi^2(1) = 32.14$, $p < .01$. There was no difference in preference due to participants' gender, $\chi^2(1) = 0.18$, $p > .5$. In Condition 2, 24 students did not indicate any preference. Thirty participants indicated a preference for a female examiner, whereas 12 participants preferred a male examiner. This difference reached significance, $\chi^2(1) = 7.71$, $p < .01$. There was no difference in preference due to participants' gender, $\chi^2 (1) = 0.36$, $p > .5$.

## 4.6 Study 2

In an online study, participants of different universities throughout Germany watched four video clips showing younger and middle-aged female and male examiners (one from each group; four videos in total) giving an introduction to a fictional intelligence test session. After having seen each clip, each person rated the examiners' characteristics in different domains. At the end, participants were required to choose a favourite examiner.

### 4.6.1 Materials and Method

**Participants.** Three hundred seventy-five students participated voluntarily and anonymously (265 women aged 19 to 54, $M = 24.4$, $SD = 4.5$, and 105 men aged 19 to 52, $M = 24.7$, $SD = 4.6$, with five participants indicating no gender). All regions of Germany were represented with a wide range of study courses covering medicine, engineering, and the social sciences.

**Materials.**

***Video clips.*** Each video clip showed the upper frontal part of the body of one examiner giving the same introduction to a test. The videos' average length was 109 seconds (*SD* = 15.85). All examiners had a degree in psychology and provided practical experience in test administration. Psychologists were four male and four female psychologists younger than 35 and four female and four male psychologists older than 44. All clips were videotaped using the same equipment, the same electrical lighting, and a white background. In the final version, each test taker evaluated one randomly selected member[2] of the different examiner groups in a randomized order of examiners.

***Expertise.*** We assessed perceived expertise of examiners with four items (e.g., "I think the examiner is proficient in her/his area of expertise"). Cronbach's alpha of this short scale ranged from .81 to .83.

***Social competence.*** We assessed perceived social competence using three items (e.g., "The examiner probably interacts well with test takers"). Cronbach's alpha ranged from .75 to .81 here.

***Social dominance orientation.*** We assessed test takers' SDO with four items adapted from the original SDO scale (Pratto et al., 1994). Internal consistency was $\alpha = .51$ in our sample.

***Preference.*** After having watched and evaluated all video clips, test takers were asked to choose whom of the four examiners they would prefer in a real test setting.

**Procedure.** Participants were contacted via e-mail. Following an online link, they were introduced to the aims and topic of the study. Each test taker then watched one video clip of each examiner group and rated each examiner directly after the presentation. Each participant then indicated an overall preference for one of the examiners, filled out the SDO items, and answered demographic questions. Finally, persons were invited to participate in a lottery. The mean time for completion of the entire survey was 25 minutes. Data were collected in the spring of 2010.

**Statistical analysis.** The following analyses were applied:

(1) To test the first hypothesis addressing systematic gender preferences, we conducted a chi-square test.

[2] One middle-aged woman refused approval of her video.

(2) To analyze stereotypic descriptions, we employed two repeated-measures ANOVAs with perceived social competence and perceived expertise as dependent variables and examiner groups (i.e., gender and age) as independent variables. To account for individual effects of the random selection of examiner videos, we included examiner ID as a factor in the model.

(3) To investigate the impact of persons' characteristics on overall preference for either male or female examiners, we conducted a loglinear analysis with participants' SDO and gender as factors. To scrutinize the impact of social competence and perceived expertise on the preference for an examiner, we employed a multinomial logistic regression analysis with social competence and expertise as covariates and random selection of examiner as factors.

### 4.6.2 Results

Referring to our first hypothesis, 248 participants (i.e., 66.13%) indicated a preference for a female examiner, whereas 127 (i.e., 33.87%) preferred a male examiner. This difference reached significance, $\chi^2(1) = 39.04$, $p < .001$. The odds of preferring a female examiner were 1.95 times the odds of preferring a male examiner.

Referring to the second research question, the first repeated measures ANOVA revealed a significant difference between the examiner groups, $F(3, 1053) = 86.39$, $p < .001$. Significant interactions between perceived social competence and the random selection of examiners were found. A separate analysis of the random selection revealed that in each examiner group, one of the examiners was rated differently from the other examiners in the group, except for the group of young women where no clear outlier was identified.[3] As expected, post hoc comparisons revealed that female examiners' social competence was rated significantly higher than male examiners' social competence. There was no significant difference between young and middle-aged women, whereas middle-aged men received significantly higher ratings than young men (see Table 4.1).

The second repeated measures ANOVA revealed a significant difference in perceived expertise, $F(3, 1041) = 70.45$, $p < .001$, between the examiner groups. Again, a significant interaction between expertise and random selection of the examiners was shown. The same individual examiners per group were rated differently on perceived expertise, except for the group of young women where no clear outlier was revealed.[3] There was a significant

[3] The differing single persons in each group had no impact on the overall significance of the main effects. Results remained the same after removing these persons.

between-subjects effect of random selection of young male examiners, $F(3, 347) = 3.94$, $p < .01$. Post hoc comparisons revealed that the expertise ratings of both male and female middle-aged examiners were significantly higher than the expertise ratings of young examiners. Furthermore, young female examiners' perceived expertise was significantly higher than young male examiners' perceived expertise (see Table 4.1).

*Table 4.1.* Means and Standard Deviations for Social Competence Ratings and Expertise Ratings

| | Examiner group | | | |
|---|---|---|---|---|
| | Young women | Young men | Middle-aged women | Middle-aged men |
| Expertise | 14.18 (3.31) | 11.97 (3.70) | 14.94 (3.40) | 14.97 (3.49) |
| Social competence | 11.93 (2.18) | 9.55 (2.49) | 11.51 (2.49) | 10.84 (2.37) |

Regarding general effects underlying examiner preferences, we employed a multinomial logistic regression testing the full model against a constant-only model. The results indicated that social competence and expertise significantly distinguish between the preference for one of the four examiner groups, $\chi^2(57) = 468.75$, $p < .01$. Nagelkerke's $R_N^2 = .80$ indicated a moderately strong relationship between prediction and grouping. For the choice of examiners, her or his rated social competence and expertise were significant predictors. The random selections of the examiners were not significant. As shown in Table 4.2, social competence was in most cases a stronger predictor than expertise for choice of each examiner group.

*Table 4.2.* Summary of Multinomial Logistic Regression Analysis for Social Competence and Expertise Predicting Preference for an Examiner Group

| Variable | | 95% CI for odds ratio | | |
|---|---|---|---|---|
| | B (SE) | Lower | Odds ratio | Upper |
| Young male examiner vs. young female examiner | | | | |
| Intercept | 10.60 (3.26) | | | |
| Young female expertise | -0.48 (0.13)*** | 0.48 | 0.62 | 0.81 |
| Young male expertise | 0.30 (0.12)* | 1.07 | 1.35 | 1.71 |
| Middle-aged female expertise | -0.12 (0.13) | 0.68 | 0.88 | 1.14 |
| Middle-aged male expertise | -0.09 (0.12) | 0.73 | 0.92 | 1.16 |
| Young female social competence | -0.89 (0.21)*** | 0.27 | 0.41 | 0.62 |
| Young male social competence | 0.58 (0.19)*** | 1.24 | 1.78 | 2.57 |
| Middle-aged female social competence | -0.13 (0.16) | 0.64 | 0.88 | 1.20 |
| Middle-aged male social competence | -0.18 (0.16) | 0.61 | 0.83 | 1.14 |
| Middle-aged female examiner vs. young female examiner | | | | |
| Intercept | -0.49 (2.05) | | | |
| Young female expertise | -0.52 (0.09)*** | 0.50 | 0.60 | 0.71 |
| Young male expertise | 0.03 (0.07) | 0.90 | 1.03 | 1.17 |
| Middle-aged female expertise | 0.45 (0.09)*** | 1.30 | 1.57 | 1.88 |
| Middle-aged male expertise | -0.02 (0.07) | 0.86 | 0.98 | 1.11 |
| Young female social competence | -0.48 (0.13)*** | 0.48 | 0.62 | 0.80 |
| Young male social competence | -0.06 (0.10) | 0.77 | 0.94 | 1.14 |
| Middle-aged female social competence | 0.70 (0.13)*** | 1.56 | 2.02 | 2.61 |
| Middle-aged male social competence | -0.13 (0.10) | 0.72 | 0.88 | 1.06 |
| Middle-aged male examiner vs. young female examiner | | | | |
| Intercept | -0.16 (2.17) | | | |
| Young female expertise | -0.41 (0.09)*** | 0.56 | 0.67 | 0.80 |
| Young male expertise | 0.14 (0.07)* | 1.00 | 1.15 | 1.32 |
| Middle-aged female expertise | -0.07 (0.07) | 0.81 | 0.94 | 1.08 |
| Middle-aged male expertise | 0.44 (0.09)*** | 1.29 | 1.55 | 1.85 |
| Young female social competence | -0.66 (0.14)*** | 0.39 | 0.52 | 0.68 |
| Young male social competence | -0.05 (0.10) | 0.78 | 0.95 | 1.16 |
| Middle-aged female social competence | -0.01 (0.11) | 0.80 | 0.99 | 1.24 |
| Middle-aged male social competence | 0.48 (0.12)*** | 1.29 | 1.62 | 2.03 |

*Note.* $R^2 = .74$ (Cox & Snell), .80 (Nagelkerke). Model $\chi^2(57) = 468.75$, $p < .01$.
*p < .05. ***p < .001.

With reference to the impact of individual differences, the loglinear analysis revealed a likelihood ratio of the final model of $\chi^2(0) = 0$, $p = 1$. Results indicated a significant interaction of SDO and gender, $\chi^2(3) = 10.37$, $p < .02$. Odds ratios showed that for female participants with high SDO, the odds of their preference for a female examiner were 1.16 times the odds of female participants with low SDO. For male participants with high SDO, the odds of their preference for a female examiner were 2.89 times the odds of male participants with low SDO. Overall, participants with higher SDO preferred female examiners (see Table 4.3).

*Table 4.3.* Percentage of Choice of Male and Female Examiners Regarding Test-Taker Gender and Test-Taker SDO

| | | Preferred examiner | |
|---|---|---|---|
| | Test-taker gender | Female | Male |
| SDO low | Woman | 22.0% | 12.2% |
| | Man | 4.6% | 4.3% |
| SDO high | Woman | 25.5% | 12.0% |
| | Man | 13.6% | 5.7% |

## 4.7 General Discussion

We conducted two studies to gain further insight into the questions of what kinds of examiner characteristics are preferred by potential test takers and how examiners are perceived. In line with our main hypothesis, both the pilot study and the online survey revealed the expected general preference for women as examiners. In the first study, this result was revealed with no further information given about potential examiners except their gender. In the second study, participants' evaluations were based on video clips of examiners introducing the test for about two minutes.

This study further revealed that – as expected – women as examiners were rated as more socially competent. Test takers indicated that they expected them to be more respectful and sensitive in a test situation. On the other hand, and opposite of our expectations, women were not evaluated as possessing less expertise than male examiners. In general, middle-aged examiners received higher ratings in expertise compared to young examiners.

We found social competence and rated expertise to be important factors; by trend social competence was a stronger predictor of preference. The preference for women as

examiners may at least in part be explained by their expected social competence. However, our data cannot explain *why* social competence is important. We cannot answer the question of whether an examiner's social competence suggests less potential ego threat, establishes a more relaxed atmosphere, or comprises both. However, both mechanisms may serve as explanations for prior results showing participants giving higher self-estimations when tested by a woman (Ortner & Vormittag, 2011).

Moreover, our study confirmed also an influence of test takers' characteristics on preferences: Participants with higher SDO indicated an even stronger preference for female examiners. This result is in line with prior research indicating that people with high SDO hold especially strong stereotypical views (Sidanius & Pratto, 2001).

As a side result, we found young men to be rated as significantly more inferior compared to examiners of all other groups: Expertise was found to be related to age (as an indicator of experience), whereas warmth seemed to be related to women (or femininity). Young men were perceived to lack both and received the most derogating evaluations. Ratings of young male psychologists in our study may have been influenced by the public discussion of young men's school underachievement (cf. Van Houtte, 2004) as well as a stereotype of young men as less warm and less sensitive to other people's needs (cf. Spence & Buckner, 2000). Examiner trainings could prepare young male examiners for stereotypic evaluations.

We are concerned with some limitations of our research: First, we did not include a real assessment, leaving open the question of whether this differential perception has a direct impact on performance. Also, our research design did not test for additional effects such as test takers' *motivated stereotyping* (Sinclair & Kunda, 2000), which could lead to devaluation of examiners due to feelings of threat or disappointment. Second, participants watched four examiners giving the same standard instructions consecutively. We decided for this artificial setting to keep differences between the administrating persons as minimal as possible, although this limited the ecological validity of our study. Third, similar to other online surveys, our study faces some methodological drawbacks as online data collection lacks monitoring and control of participants' behaviour. Although it is difficult to gain sufficient data in real test situations, future studies may include perceptions in evaluations of assessment procedures. Moreover, further studies may include a behavioural analysis of examiners in order to analyse whether behavioural differences between men and women underly the effects we found.

Further research in real assessment procedures is therefore needed. Additionally, future studies should include more examiners to explore the impact of further and additional person characteristics. For example, research has indicated that perceived examiners' attractiveness may serve as a further variable (Karremans et al., 2009).

The present study is the first to show a preference for women as examiners in the setting of aptitude and achievement tests. Our results reveal that even when using standardized assessment procedures, examiner characteristics impact test takers' perceptions and expectations with regard to the test situation. If perceptions and preferences lead to differences in de facto performance (e.g., benefiting from a female examiner), this would violate the basic claim of fairness in psychological assessment, demanding equal opportunities to show maximum performance for all test takers (see APA, AERA, & NCME, 1999). It is possible then that providing a choice of examiner in individual assessments and interviews may present a solution. Furthermore, computerized assessment or having several male and female examiners present may minimize an individual examiner's impact. Future studies could investigate how preferences for certain examiners influence performance under the preferred versus not preferred examiners, and studies could be conducted on examiner preferences and their consequences in further domains such as clinical assessment, for example.

## 4.8 References

Allport, F. H. (1954). The Structuring of Events - Outline of a General Theory with Applications to Psychology. *Psychological Review, 61*, 281-303. doi: 10.1037/h0062678

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*, 656-665. doi: 10.1037/0022-0663.87.4.656

Basow, S. A. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles, 43*, 407-417. doi: 10.1023/A:1026655528055

Danso, H. A., & Esses, V. M. (2001). Black experimenters and the intellectual test performance of white participants: The tables are turned. *Journal of Experimental Social Psychology, 37*, 158-165. doi: 10.1006/jesp.2000.1444

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin, 26*, 1171-1188. doi: 10.1177/0146167200262001

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles, 47*, 99-114. doi: 10.1023/A:1021020920715

European Commission (Ed.). (2011). *Report on Progress on Equality between Women and Men in 2010 - The gender balance in business leadership*. Luxembourg: Publications Office of the European Union, 2011. doi: 10.2767/99441

Fernández-Ballesteros, R. (1999). Psychological assessment: Future challenges and progresses. *European Psychologist, 4*, 248-262. doi: 10.1027//1016-9040.4.4.248

Fernández-Ballesteros, R., De Bruyn, E. E. J., Godoy, A., Hornke, L. F., Ter Laak, J., Vizcarro, C., et al. (2001). Guidelines for the Assessment Process (GAP): A proposal for discussion. *European Journal of Psychological Assessment, 17*, 187-200. doi: 10.1027//1015-5759.17.3.187

Fiske, S. T., & Neuberg, S. L. (1990). A Continuum of Impression-Formation, from Category-Based to Individuating Processes - Influences of Information and Motivation on Attention and Interpretation. *Advances in Experimental Social Psychology, 23*, 1-74.

Fiske, S. T., & Taylor, S. E. (1996). *Social Cognition* (2nd ed.). New York: McGraw Hill.

Glick, P., Larsen, S., Johnson, C., & Branstiter, H. (2005). Evaluations of sexy women in low- and high-status jobs. *Psychology of Women Quarterly, 29*, 389-395. doi: 10.1111/j.1471-6402.2005.00238.x

Harris, M. J., & Rosenthal, R. (1985). Mediation of Interpersonal Expectancy Effects - 31 Meta-Analyses. *Psychological Bulletin, 97*, 363-386. doi: 10.1037/0033-2909.97.3.363

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47*, 237-271. doi: 10.1146/annurev.psych.47.1.237

Huang, M. H. (2009). Race of the interviewer and the black-white test score gap. *Social Science Research, 38*, 31-40. doi: 10.1016/j.ssresearch.2008.07.004

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo. *Political Psychology, 25*, 881-919. doi: 10.1111/j.1467-9221.2004.00402.x

Karremans, J. C., Verwijmeren, T., Pronk, T. M., & Reitsma, M. (2009). Interacting with women can impair men's cognitive functioning. *Journal of Experimental Social Psychology, 45*, 1041-1044. doi: 10.1016/j.jesp.2009.05.004

Kite, M. E., Stockdale, G. D., Whitley, B. E., & Johnson, B. T. (2005). Attitudes toward younger and older adults: An updated meta-analytic review. *Journal of Social Issues, 61*, 241-266. doi: 10.1111/j.1540-4560.2005.00404.x

Ko, S. J., Judd, C. M., & Stapel, D. A. (2009). Stereotyping Based on Voice in the Presence of Individuating Information: Vocal Femininity Affects Perceived Competence but Not Warmth. *Personality and Social Psychology Bulletin, 35*, 198-211. doi: 10.1177/0146167208326477

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*, 284-308. doi: 10.1037/0033-295X.103.2.284

Mishra, S. P. (1980). Influence of Examiners Ethnic Attributes on Intelligence-Test Scores. *Psychology in the Schools, 17*, 117-122. doi: 10.1002/1520-6807(198001)17:1<117::AID-PITS2310170122>3.0.CO;2-6

Muniz, J., & Bartram, D. (2007). Improving international tests and testing. *European Psychologist, 12*, 206-219. doi: 10.1027/1016-9040.12.3.206

Ortner, T. M., & Vormittag, I. (2011). Test administrator's gender affects female and male students' self-estimated verbal general knowledge. *Learning and Instruction, 21*, 14-21. doi: 10.1016/j.learninstruc.2009.09.003

Posthuma, R. A., & Campion, M. A. (2009). Age stereotypes in the workplace: Common stereotypes, moderators, and future research directions. *Journal of Management, 35*, 158-188. doi: 10.1177/0149206308318617

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social-Dominance Orientation - a Personality Variable Predicting Social and Political-Attitudes. *Journal of Personality and Social Psychology, 67*, 741-763. doi: 10.1037/0022-3514.67.4.741

Redman, T., & Snape, E. (2002). Ageism in teaching: stereotypical beliefs and discriminatory attitudes towards the over-50s. *Work Employment and Society, 16*, 355-371. doi: 10.1177/095001702400426884

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research*. New York, N.Y.: Irvington Publishers, Inc.

Sidanius, J. & Pratto, F. (2001). Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression. Cambridge: University Press

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin, 26*, 1329-1342. doi: 10.1177/0146167200263002

Spence, J. T., & Buckner, C. E. (2000). Instrumental and expressive traits, trait stereotypes, and sexist attitudes - What do they signify? *Psychology of Women Quarterly, 24*, 44-62. doi: 10.1111/j.1471-6402.2000.tb01021.x

Spencer, S. J., Fein, S., Wolfe, C. T., Fong, C., & Dunn, M. A. (1998). Automatic activation of stereotypes: The role of self-image threat. *Personality and Social Psychology Bulletin, 24*, 1139-1152. doi: 10.1177/01461672982411001

Taylor, S. E., & Lobel, M. (1989). Social-Comparison Activity under Threat - Downward Evaluation and Upward Contacts. *Psychological Review, 96*, 569-575. doi: 10.1037/0033-295X.96.4.569

Van Houtte, M. (2004). Why boys achieve less at school than girls: The difference between boys' and girls' academic culture. *Educational Studies*, 30, 159-173. doi: 10.1080/0305569032000159804

Westhoff, K., Hagemeister, C., Kersting, M., Lang, F., Moosbrugger, H., Reimann, G., et al. (2010; 3rd ed.) *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430*. Lengerich: Pabst Science Publisher.

Wills, T. A. (1981). Downward Comparison Principles in Social-Psychology. *Psychological Bulletin, 90*, 245-271. doi: 10.1037/0033-2909.90.2.245

# Chapter 5

# Better Cognitive Functioning in the Presence of Women! Situational Influence on Test Performance in Large-Scale Assessment

Vormittag, I. & Ortner, T. M. (submitted). Better Cognitive Functioning in the Presence of Women! Situational Influence on Test Performance in Large-Scale Assessment.

## 5.1 Abstract

In modern western societies testing and psychological assessment have become an inherent part of educational and vocational processes. Prior research indicated that even in standardized assessment the performance of test takers may be influenced by the examiner. In this study we used the data of 2,862 participants between the ages of 17 and 70 of a large-scale survey taking a short speed test to investigate possible performance differences due to examiner gender. The test was applied on laptops with one of 178 examiners (86 women and 92 men) present during the testing. We applied multilevel analyses and found a small but significant effect of examiner gender: Test takers performed better with a female examiner present. Effect sizes are reported and possible explanations for the results discussed.

*Keywords:* Examiner effect; Large-scale assessment; Cognitive performance

## Better Cognitive Functioning in the Presence of Women!
## Situational Influence on Test Performance in Large-Scale Assessment

In modern western societies, people typically undergo a number of testing and examination situations throughout their lives in grade school, at college, and in the work context (e.g., Fernández-Ballesteros, 1999). In oral-examination situations, paper–pencil testing, and even in computerized assessment, there is usually one but sometimes several people involved as examiners (Anastasi & Urbina, 1997). It is an everyday observation that people often state examiner preferences indicating that test and examination settings are not simply situations of cognitive load, but also contain a social component.

In fact, early research revealed that test takers' cognitive performances systematically differ according to examiner characteristics in experiments or test situations (cf. Rosenthal, 1976; Rosenthal & Fode, 1963). In addition to effects that were found to be related to examiners' ethnic attributes (Danso & Esses, 2001; Katz, Robert, & Robinson, 1965), an early review summarized overall mixed results with reference to examiner's gender (see Rumenik, Capasso, & Hendrick, 1977).

The knowledge base has not changed considerably since then (Ortner & Vormittag, 2011). At the present time, there is a general lack of research on the possible impact of examiner's gender on intellectual performance. Nevertheless, the question of examiner effects could be addressed by employing data using contemporary large-scale surveys. Such surveys offer the possibility of including data not only with regard to different age groups and educational levels of participants, but also with regard to different interviewers or examiners. With respect to self-report data, studies have already revealed examiner and interviewer effects in large-scale assessments (e.g., Kish, 1962; Schnell & Kreuter, 2005). Catania et al. (1996) addressed self report on sexual behavior and found that the majority of women chose a female interviewer, whereas the choice of male participants was more balanced with a trend to preferred female interviewers. Groves and Fultz (1985) compared male and female interviewers of a large survey centre and found higher response rates for female interviewers, but no differences in missing data of participant's interviews. They revealed that on questions concerning the economic future respondents answered more optimistically when interviewed by a man. These different answer patterns emerged independently of respondent's gender. The authors assumed different perception of male and female interviewers as possible explanation. Only one recent study addressed performance in large-scale assessment and revealed effects of examiner's ethnicity (Huang, 2009).

The aim of the present study was, for the first time, to more extensively respond to the question of the possible impact of examiner's gender on test takers' cognitive performance by employing a representative, large-scale data set from the German Socio-Economic Panel (GSOEP).

## 5.2 Method

GSOEP is a wide-ranging representative longitudinal study of private households. The data provides information on all household members, consisting of Germans, foreigners, and recent immigrants. Data collection started in 1984. Some of the many topics include household composition, occupational biographies, employment, earnings, health, and satisfaction indicators. In the year 2006, the first time more than 5,500 participants of the GSOEP were invited to participate in a short achievement testing after the interview including a perceptual speed task.

### 5.2.1 Participants

**Test takers**. In total, 1,501 women and 1,362 men between the ages of 17 and 70 ($M$ = 44.55, $SD$ = 14.21) were included. We excluded persons who needed the examiner's assistance due to physical impairments.

**Examiners**. In the original data set examiners' age ranged from 26 to 70 years, with the majority of examiners being older than 50. We excluded data collected by examiners older than 65 years in order to keep the results comparable with research on examiner effects in organizational or educational context. We furthermore used only data of examiners older than 44, because younger examiners represented less than 5 percent of the data. Based on this procedure, data gained by 178 examiners (86 women and 92 men) between the ages of 44 and 65 (M = 56.52, SD = 5.83) was used. Some examiners had already interviewed some participants in previous years and were present during testing.

### 5.2.2 Materials

**Cognitive performance test**. A short speed test was employed requesting participants to assign numbers to graphical symbols based on given rules (see Schupp, Herrmann, Jaensch, & Lang, 2008) on a laptop. Total time for the testing phase was 90 seconds. Different variables were obtained with the total score of numbers within 90 seconds as the main outcome variable.

### 5.2.3 Analysis

We conducted multilevel analyses allowing us to control for the random influence of individual examiners (e.g. Raudenbush & Byrk, 2002). Examiner's gender (dummy coded: 0 = female, 1 = male; Level 2) and participant's age (Level 1) were included as predictors in a random intercept and slope model, as research has found increasing age to be related to lowered perceptual speed (Salthouse, 1994). The first model was an intercept-only model; the second included participant's age as a predictor (Level 1 model), and the third additionally included examiner's gender (full model).

## 5.3 Results

The final random intercept and slope model explained the data significantly better than the first intercept-only model, $\chi^2(4) = 592.01$, $p < .05$, and the model with only participants' age as a predictor (Level 1 model), $\chi^2(1) = 4.25$, $p < .05$ (see Table 5.1).

*Table 5.1.* Summary of the Three Models of Multilevel Analyses Predicting Performance in Perceptual Speed

| | Intercept-only model | Level 1 model | Full model |
|---|---|---|---|
| Fixed effects (parameter estimates) | | | |
| Intercept | 28.32*** | 28.25*** | 29.27*** |
| Participant age (centered) | | -3.67*** | -3.67*** |
| Examiner gender | | | - 1.90 * |
| Random effects (covariance parameter estimates) | | | |
| Variance in means | 32.50*** | 30.06*** | 29.35*** |
| Covariance between means and slopes | | - 3.51* | - 3.70* |
| Variance in slopes | | 2.25* | 3.10* |
| Variance within groups | 71.56*** | 56.72*** | 56.71*** |
| -2 log likelihood (df) | 20648.58 (3) | 20060.82 (6) | 20056.57 (7) |

*Note. ICC = .31. $R1^2$ = .21, $R2^2$ = .10.*
**p < .05. ***p < .001.*

Parameter estimates showed that the age of participants was negatively correlated with performance, $b$ = -3.67, $t(96.05)$ = -16.99, $p < .001$. Furthermore, participants performed better when the test was presented in the presence of a female examiner, $b$ = -1.90, $t(151.82)$ = -2.08, $p < .05$. Finally, 10% of performance variance on the examiner level (Level 2) was explained by examiner's gender. Descriptive results demonstrate that participants tested by a woman achieved a higher mean score than participants tested by a man (see Table 5.2).

*Table 5.2.* Means and Standard Deviations of Test Scores

| | Examiners | |
|---|---|---|
| Participants | Women | Men |
| Women | 29.03 (9.46) | 27.38 (9.99) |
| Men | 29.02 (10.07) | 27.55 (10.36) |

## 5.4 Discussion

The present study is the first that aimed to investigate the impact of examiner's gender on test taker's performance on a standardized achievement test using a large data sample. The two-level hierarchical model based on a large representative sample revealed that besides the influence of test takers' age that is well in line with previous research (Salthouse, 1994), examiner's gender had a small but significant impact on participant's performance in a speed test. The result is especially notable because the test was not presented in a face-to-face testing situation, but in computerized form with merely the presence of the examiner. Examiner's gender influenced the performance although persons worked with the computer on their own. Due to lacking data covering all examiners' age groups sufficiently, we restricted the age to examiners between 44 years to 65 years. This could have leveled possible age effects. However, the age range still covered more than twenty years. Therefore our results apply at least to middle-aged examiners.

Different explanations could be proposed for the significant influence of the examiner's gender: First, it could be assumed that male and female examiners behave differently. Although standardization is requested, the possibility remains that examiners, for example, were differently motivating. Past research in experimental psychology has revealed that in fact, men and women differ in their nonverbal communication behavior (Hall, Coats, &

Smith LeBeau, 2005; Steckler & Rosenthal, 1985). Unconscious, nonverbal cues have shown to influence participant's performance (Rosenthal, 1976; 2002).

The data collection of the GSOEP was not designed to answer such questions; therefore possible behavioral differences of male and female examiners cannot be investigated here. Future studies on examiner gender effects may therefore analyze behavioral differences.

Second, stereotypes conforming to the perception of female examiners may serve as an explanation: Women could have been evaluated as more sensitive and as creating a more positive or relaxed atmosphere (Samuel, 1977; Spence & Buckner, 2000). By contrast, the testing situation could have been perceived as more ego-threatening with a male examiner who may be seen as more competent or intelligent (Beloff, 1992; Ortner, Müller, & Garcia-Retamero, 2011). However, we cannot conclude that the effect we found was a result of test takers' perceptions and interpretations, examiners' behaviors, or a mixture of both.

Our investigation faces several limitations: First, the interview setting of the GSOEP is not in accordance with a standardized testing situation. Although interviewers receive prior training and are requested to establish a concentrated atmosphere, testing still took place in a private household where control of the testing environment is limited. Second, we used data on only one achievement related test assessing a single aspect of cognitive ability. Therefore we cannot conclude if results generalize to other achievement domains. Third, the test takers participated voluntarily. Results therefore cannot be transferred to other situations where assessment procedures are obliged and/or results are associated with relevant consequences. Finally, the difference we found can be classified as a small effect (Cohen's *d* for the performance difference between female and male examiners was 0.17 for women and 0.14 for men). As Willingham and Cole (1997) pointed out, even very small group differences may produce great factual effects in large-scale assessment procedures if only few persons are selected. However, the small effect may to some degree serve as an explanation for the previous ambiguous results that were found (Rumenik, Capasso, & Hendrick, 1977).

As contemporary approaches that aim to enhance the quality of decisions based on psychological tests and standardized questionnaires mostly address more *technical* aspects of the measures applied, our study indicates that future efforts should also address situational aspects in order to increase objectivity and fairness (e.g., Fernández-Ballesteros, 1999).

## 5.5 References

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River (NJ): Prentice Hall.

Beloff, H. (1992). Mother, father and me: Our IQ. *The Psychologist, 5*, 309-311.

Catania, J. A., Binson, D., Canchola, J., Pollack, L. M., Hauck, W., & Coates, T. J. (1996). Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior. *Public Opinion Quarterly, 60*, 345-375. doi: 10.1086/297758

Danso, H. A., & Esses, V. M. (2001). Black experimenters and the intellectual test performance of white participants: The tables are turned. *Journal of Experimental Social Psychology, 37*, 158-165. doi: 10.1006/jesp.2000.1444

Fernández-Ballesteros, R. (1999). Psychological assessment: Future challenges and progresses. *European Psychologist, 4*, 248-262. doi: 10.1027//1016-9040.4.4.248

Groves, R. M., & Fultz, N. H. (1985). Gender Effects among Telephone Interviewers in a Survey of Economic Attitudes. *Sociological Methods & Research, 14*, 31-52. doi: 10.1177/0049124185014001002

Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin, 131*, 898-924. doi: 10.1037/0033-2909.131.6.898

Huang, M. H. (2009). Race of the interviewer and the black-white test score gap. *Social Science Research, 38*, 31-40. doi: 10.1016/j.ssresearch.2008.07.004

Katz, I., Roberts, S. O., & Robinson, J. M. (1965). Effect of difficulty, race of administrator, and instructions on Negro digit-symbol performance. *Journal of Personality and Social Psychology, 2*, 53-59. doi: 10.1037/h0022080

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. *Journal of the American Statistical Association, 57*, 92-115.

Ortner, T. M., Müller, S. M., & Garcia-Retamero, R. (2011). Estimations of parental and self intelligence as a function of parents' status: A cross-cultural study in Germany and Spain. *Social Science Research, 40*, 1067-1077. doi: 10.1016/j.ssresearch.2011.03.006

Ortner, T. M., & Vormittag, I. (2011). Test administrator's gender affects female and male students' self-estimated verbal general knowledge. *Learning and Instruction, 21*, 14-21. doi: 10.1016/j.learninstruc.2009.09.003

Raudenbush, S. W., & Byrk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.

Rosenthal, R. (1976). Experimenter Effects in Behavioral Research (enlarged edition). New York: Irvington Publishers.

Rosenthal, R. (2002). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist, 57*, 839-849. doi: 10.1037/0003-066X.57.11.839

Rosenthal, R., & Fode, K. L. (1963). Three experiments in experimenter bias. *Psychological Reports, 12,* 491-511.

Rumenik, D. K., Capasso, D. R., & Hendrick, C. (1977). Experimenter Sex Effects in Behavioral Research. *Psychological Bulletin, 84*, 852-877. doi: 10.1037/0033-2909.84.5.852

Salthouse, T. A. (1994). The Nature of the Influence of Speed on Adult Age-Differences in Cognition. *Developmental Psychology, 30*, 240-259. doi: 0012-1649/94/S3.00

Samuel, W. (1977). Observed IQ as a function of test atmosphere, tester expectation, and race of tester: A replication for female subjects. *Journal of Educational Psychology, 69*, 593-604. doi: 10.1037/0022-0663.69.5.593

Schnell, R., & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics, 21*, 389-410.

Schupp, J., Herrmann, S., Jaensch, P., & Lang, F. R. (2008). Erfassung kognitiver Leistungspotentiale Erwachsener im Sozio-oekonomischen Panel (SOEP) [Surveying Cognitive Potentials of Adults in the Socio-Economic Panel (SOEP)]. Berlin: Deutsches Institut für Wirtschaftsforschung.

Socio-Economic Panel (SOEP) data for the years 1984-2009, version 26. (2010).

Spence, J. T., & Buckner, C. E. (2000). Instrumental and expressive traits, trait stereotypes, and sexist attitudes - What do they signify? *Psychology of Women Quarterly, 24*, 44-62. doi: 10.1111/j.1471-6402.2000.tb01021.x

Steckler, N. A., & Rosenthal, R. (1985). Sex-Differences in Nonverbal and Verbal Communication with Bosses, Peers, and Subordinates. *Journal of Applied Psychology, 70*, 157-163. doi: 10.1037/0021-9010.70.1.157

Willingham, W. W., & Cole, N. S. (1997). Introduction. In W. W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 1-15). Mahwah, NJ: Erlbaum.

# Chapter 6

# General Discussion

# General Discussion

## 6.1 Summary of the Results

This doctoral thesis covered four studies to investigate examiner effects on test takers' performance in standardized tests with special regard to examiner's gender. Table 6.1 gives an overview of the specific aims, methods, and findings of the four studies.

*Table 6.1.* Summary of aims, methods, and major findings of this thesis

| Chapter | Aim | Methods | Findings |
|---|---|---|---|
| 2 | Investigation of effects of examiners' gender in a standardized setting. | In a quasi-experimental research design a verbal knowledge test was employed. Analyses of variance were applied for group comparison. | Test takers estimated their verbal knowledge higher when tested by a female examiner. |
| 3 | Increase of explained variance of gender effects by including an additional examiners' characteristic | In a quasi-experimental research design a verbal knowledge test and ratings on examiner attractiveness were used. Moderated regression analyses were employed. | Self-estimations in presence of examiners rated as attractive were lowered. In same-gender interaction attractiveness of the examiner worsened performance. |
| 4 | Investigation of preference for female or male examiners and exploration of perceptions of examiners. | In a pilot study participants chose a specific examiner. In the main study video clips of examiners were rated online. For analyses chi-square tests, analyses of variance, log-linear analysis, and logistic regression were used. | In both studies women were preferred as examiners. Middle-aged examiners were perceived as more proficient independent of gender. Independent of age, women were rated as more socially competent. |
| 5 | Investigation of performance differences due to examiner gender in a large sample. | Performance on a short speed test in a large survey (GSOEP) was analyzed. Multilevel analyses were employed. | Participants performed significantly better when tested by a woman, although effect sizes were small. |

*Note. GSOEP = German Socio-economic panel*

In this chapter the contributions of the single studies to the main goals of this thesis are discussed, congruent and contradictory aspects are regarded. Implications for the testing practice are presented. This chapter closes with final concluding remarks.

### 6.1.1 Results Gained on Effects due to Examiners' Gender in Standardized Testing

The primary goal of this thesis was to investigate if examiners' gender influences test takers in standardized tests. In Study 1 (Chapter 2) and Study 4 (Chapter 5) of this thesis the occurrence of effects of examiners' gender were revealed on test takers' intellectual performance. Furthermore, one study (Study 1) revealed an effect of examiners' gender on test takers' self-estimated knowledge. In Study 1 examiner gender effects were shown in a face-to-face test setting employing a verbal knowledge test. Two scores were obtained: a subjective score – i.e. self-estimated verbal knowledge – and an objective score – namely the number of correctly solved items in the verbal knowledge test. It was assumed, that the comparison of subjective scores would evoke greater effects than the comparison of objective scores. Furthermore, the subjective score was used as an indicator of a metacognitive experience during task completion (Efklides, Samara et al. 1999; Efklides 2006). Test takers estimated their own verbal knowledge more positive when tested by a female examiner. Effect size for differences in self-estimated knowledge was almost medium (Cohen's $d$ = .46) referring to Cohen's (1992) interpretation of effect sizes. Concerning the de facto knowledge, no significant difference emerged. Nevertheless, results indicated a small trend towards better performance under female administration.

In Study 4 examiner effects were also revealed for participants in a large survey who worked on a computerized speed test with one examiner present during the testing. Test takers performed better in presence of a female examiner on this speed test. The differences in performance mean for participants with male versus female examiner were small (Cohen's $d$ = .17 for female participants and Cohen's $d$ = .14 for male participants) and explained variance on the examiner level was only 10%. The examiner effect in this context is of special relevance though, because no direct interaction between examiner and test taker during the testing occurred. It was the first study revealing examiner gender effects on test taker's cognitive performance in a large survey sample.

In sum, both studies, Study 1 and Study 4, have strengths and weaknesses that may complement their informative value: The examiner effect revealed in the data of the GSOEP in Study 4 has been characterized by a size that would have remained insignificant in smaller samples. An estimation of the requested sample size (GPower, Faul, Erdfelder, Buchner, &

Lang, 2009) to reveal such small effect sizes with at least power of 80% indicated more than 1,000 participants. This underlines the major advantage of test data gained from a large-survey context. The result indicating small effect sizes may also serve as an explanation for the inconsistencies of previous results: In studies reported by Rumenik, Capasso, and Hendrick (1977) investigating examiner gender effects on tests, sample sizes never exceeded 500. Also other studies exploring examiner effects stayed below this sample size mark (e.g., Bookout & Hosford, 1969; Samuel, 1977). However, there are also disadvantages when employing the large survey approach that have to be taken into consideration. For example, the fact that for GSOEP trained lay persons interviewed participants in their homes. It may be proposed that the survey character may also have reduced the informative value for conclusions regarding test practice, as the survey setting does not fully assure standardized assessment. However, Study 1 addressed this issue by establishing a standardized test setting. It may be concluded that the very small impact of examiner gender on test takers de facto verbal knowledge – characterized by no significant mean differences – has been based on the sample size which was considerably smaller that the sample in Study 4. For Study 1, estimation of requested sample size to reveal these mean differences in performance with an acceptable power of 80% resulted in a required sample of 708 persons. Indeed, conducting such an experimental face-to-face study for scientific purpose is almost impossible within most university settings. Summarizing, from the results of Study 1 and Study 4 it may be concluded that during a face-to-face standardized testing not only a test taker's cognitive performance may be affected by the examiner's gender but also the self-estimation of the test taker.

Study 2 (Chapter 3) employed the same verbal knowledge test in a face-to-face test setting as Study 1 and additionally included test taker's ratings of examiner's attractiveness. In this study no main effect of examiner's gender emerged. Instead, with reference to self-estimated knowledge, perceived examiners' attractiveness led to more cautious self-estimations of both female and male test takers. This result explained 6% of variance, implying a small effect size for this main effect of perceived attractiveness ($f^2 = .06$). With reference to de facto knowledge a three way interaction was revealed: Female test takers performed worse when tested by an attractive female examiner compared with female test takers tested by a woman or a man perceived as less attractive. The same pattern – even stronger – emerged for male test takers. Explained variance of this final model was 19% with a medium effect size ($f^2 = .16$) and a power of 96%. The results of Study 2 suggest that perceived attractiveness has a strong impact on social interaction – even in a standardized test

situation. The study revealed for the first time the potential impact of attractiveness for intellectual performance in a standardized face-to-face test setting. Only one study by Karremans, Verwijmeren, Pronk, and Reitsma (2009) dealt with a similar question and revealed a negative impact of female experimenter attractiveness on a subsequent computer test result of men. The authors explained the results with increased impression management processes of male test takers when confronted with an attractive female experimenter. However, the test situations of the two studies differed: In the study of Karremans et al. the test was computerized and the female experimenter had a conversation with the test taker during an apparent test break, whereas in the study of this thesis the testing was conducted orally in a face-to-face setting. As the results of the thesis contradict those by Karremans et al. the consistence of the assumed processes has to be further investigated.

Study 1 and Study 2 revealed different results with reference to examiner gender effects. Differences in self-estimated knowledge due to examiner gender were revealed in Study 1, but not in Study 2. Two possible explanations for these diverging results should be considered: First, due to the smaller effect sizes random effects as explanations cannot be fully ruled out. Second, one main difference in the procedure may be explained by examiner samples: In the first study all of the examiners were diploma students on master student level. In the second study a major part – i.e. almost twice as much – of the examiners were bachelor students. Although mean age of examiners did not differ significantly between the two studies, a thorough view on the individual testing dyads revealed that in the first study the age gap between examiner and test taker was mostly larger than in the second study. This age difference results in a small effect size of Cohen's $d = .26$ between the two data sets. It could be assumed that age is positively associated with perceived expertise. Test takers interacting with an older examiner could have ascribed this examiner a stronger expert status than test takers in interaction with an examiner seemingly the same age as themselves. Furthermore, one could propose that due to differences in academic and field experiences the bachelor students – although instructed identically – may behaved less proficient and formal compared with the diploma students. In line with these assumptions it could be proposed that test takers perceived the examiners in the first study as more competent compared to test takers of the second study seeing examiners in the setting more or less as peers. This may have elicited other interpersonal processes in Study 2. In section 6.1.2 possible underlying attributions that may help explaining the results are proposed. However, it has to be regarded, that with the obtained data these suggestions cannot be scrutinized.

With reference to occurrence, size and direction of examiners' gender effects I come to the following conclusions: All together the results of this thesis are not fully consistent regarding occurrence of systematic examiners' gender differences. In two out of three studies examiner's gender influenced test taker's cognitive performance – one time significantly and one time insignificantly. Moreover in two studies test taker's self-estimations were assessed: In one of these studies the examiner gender had a significant impact on the test taker's self-estimation. Referring, to the first research aim, I conclude that examiner gender has a small impact on cognitive performance of test takers in face-to-face test settings. Concerning effect sizes, the first and the fourth study indicate that large sample sizes may be necessary to reveal such examiner gender influence. Furthermore, examiner gender may have an influence on self-estimated knowledge as well, although the results of these studies are not fully conclusive. Referring to the direction of effects, results are more or less consistent: In general positive ramifications of female examiners were shown in this thesis.

However, limitations of those three studies imply different claims for future research. Especially four aspects seem to be of major relevance and will be described in the following:

First, this thesis only comprises individual testings. Future studies are needed to investigate examiner effects in group testing. On the one hand, the group itself could enlarge or minimize examiner effects. Former research revealed the impact of group composition on the emergence of *stereotype threat* (e.g., Inzlicht & Ben-Zeev, 2000; Sekaquaptewa & Thompson, 2003). It seems possible that the small effects of examiner's gender may accumulate with other effects in a group testing. On the other hand, group testing could protect against examiner effects. One could argue that the influence of a single examiner is reduced in front of a group of test takers due to restricted individual interaction. Therefore research investigating group settings with regard to different group composition and size are needed. One more recent study by Lüdtke, Robitzsch, Trautwein, Kreuter, and Ihme (2007) addressed examiner effects in large-scale group testing on the results of the Third International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA) in group settings. No differences in mathematics performance were explained by the examiner's gender or experience here. This result may be interpreted as a first sign that that group settings are more protective with reference to examiner effects than single face-to-face assessment.

Second, a common feature of the discussed studies is the voluntariness of participation. This limits the generalizability of the results to real life assessment where test results have important consequences for the test taker. Although the test situations were

realistic at least in the first and second study and probably included the possibility of ego threat, it is possible that test taker's achievement motivation would have been different in real high stakes testing.

Third, the used data refers to a test on verbal knowledge and a test on perceptual speed. Future research should explore examiner effects on performance in different tests assessing different aspects of intelligence or knowledge. It should be investigated if the revealed effects only occur with specific test content, increase with more threatening test material, or vary with test length.

Finally, none of the studies analyzed examiners' behavior. Therefore, based on the obtained results of this thesis differences in examiner behavior cannot be precluded. Experimental research has shown that examiners may influence participants via nonverbal behavior (cf. Rosenthal, 1976). Although blatant behavioral differences can be diminished via standardization (cf. Sattler & Theye, 1967), the occurrence of unobtrusive gestures, differences in facial expression, or differences in intonation cannot be ruled out. For example, gender differences have been found in smiling (LaFrance, Hecht, & Levy Paluck, 2003) and nonverbal communication styles (Hall, Coats, & Smith LeBeau, 2005; Steckler & Rosenthal, 1985). Expressive, enforcing nonverbal behavior could motivate the test taker striving for best performance. Contrary, it seems possible that nonverbal communication can be distracting and irritate the test taker. Future research focusing on nonverbal behavior of the examiner is requested.

### 6.1.2 Results Gained on Attributions on Examiners in Face-to-face Testing

Besides the investigation of occurrence, size and direction of examiners' gender effects on test takers' self-estimation and performance in standardized testing, this thesis aimed to obtain information concerning underlying attributional processes on the examiners that may help explaining the emergence of examiner effects. Two studies of this thesis addressed this aspect: Study 3 (Chapter 4) laid emphasis on examiner preferences and how examiners were perceived. As described above, Study 2 included attractiveness as one additional aspect. In the following, I will outline the main results of these studies with reference to attributional processes.

Study 3 aimed to provide insight into examiner preferences and how examiners were perceived both in a real-life setting and employing an online survey approach. First, the pilot study and the main study accordingly were the first studies so far revealing that a higher amount of persons preferred female examiners compared to male examiners. Both studies

differed with reference to important aspects: First, in the pilot study, persons anticipated a cognitive testing, whereas in the online survey persons were aware that no real testing would take part. Second, in the pilot study participants had no information about the suggested two examiners apart from their gender (and hypothetical names), whereas in the online survey, examiners were presented on videos. Lack of information in the pilot study may serve as one reason for the large amount of indecisive participants. However, those participants who made a decision significantly more often decided for the female examiner. In the online survey, again the majority of persons chose a female examiner – age seemed to be of minor relevance here.

The following results were revealed in this thesis regarding the question, how examiners were perceived: Study 3 indicated significant group differences in ascribed characteristics regarding social competences and expertise, although examiners provided the identical standardized information and holding the same position in front of the camera. Middle-aged examiners (older than 44 years) received higher ratings on expertise compared with young examiners (younger than 35 years) and female examiners were rated as more socially competent than male examiners. Comparing the mean expertise ratings of middle-aged examiners with those of young examiners revealed a medium effect size (Cohen's *d* of .54). Mean ascribed differences between female and male examiners in social competence were also of medium size with Cohen's $d = .64$. Female examiners of both age groups were perceived as more socially competent. This is in accordance with common gender stereotypes ascribing women more social, sensitive, and warm traits as men (e.g., Spence & Buckner, 2000). Middle-aged examiners more expertise was ascribed. Probably, perceived higher age was interpreted as more experience and knowledge. Therefore, the results suggest, that – given the fact, that examiners behaved highly similarly – stereotypes have influenced how examiners were perceived.

Unexpectedly, perceived expertise was not related to gender in this study. Middle-aged women received high expertise ratings. On the one hand, this result may indicate, that professional psychologists expertise is ascribed independent from their gender. This explanation could be supported by a more feminine image of Psychology as a profession (Olos & Hoff, 2006). On the other hand, the results may be explained in line with the *shifting standards model* (Biernat & Kobrynowicz, 1997; Biernat & Manis, 1994). In this model Biernat describes that people requested to rate targets on Likert-type scales (e.g. "very competent" or "not competent") use a stereotype based frame of reference. Therefore, Likert-type scales are subjective measures compared to objective measures where participants are

requested to indicate percentages or rank orders. Following Biernat's reasoning persons may employ different frames of reference when rating women versus men: When asked to rate the expertise of a female examiner, participants may have implicitly compared this female examiner with other women. On the other hand, when asked to rate the expertise of a male examiner the participants may have applied a different frame of reference – namely that of men. Based on the assumption that women often are perceived as less proficient or competent, this would suggest that a lower minimum standard for female examiners may have been used. In other words, a female examiner behaving highly similarly to male examiners may appear competent for a woman. Therefore, equal or even higher ratings in proficiency for women could be explained. Nevertheless, the given data does not allow a test of this assumption.

A surprising side result concerned young male examiners: Young men were significantly less favored as examiners and received lower ratings on expertise and social competence compared to the other examiner groups. It may be suggested that higher age was associated with expertise and femininity with social competence, which led to low ratings for young men on both scales. Applying the shifting standards model once more (Biernat & Manis, 1994), one reason may be drawn from the possibility that young male examiners were compared with a high standard of competent men. Low ratings on social competence cannot be explained with this model though.

Concerning the second study several reasons may be considered when interpreting the described results: As referred above, similar age of examinees and examiners in Study 2 may have created a peer setting atmosphere with an increased facilitation of automatic social comparison processes. In general, confrontation with someone attractive is known to possibly lead to contrast effects in terms of more negative self perception (e.g., Thornton & Moore, 1993). This could explain the lowered self-estimated knowledge in interaction with an attractive examiner.

In line with this the same-gender effect on de facto knowledge may also be explained: Test takers performed worse when tested by an attractive same-gender examiner. Automatic social comparisons are even more facilitated when confronted with someone similar concerning for example gender, age, or status (Festinger, 1954; Campbell & Tesser, 1985). The fact that test taker's performance was worsened only in the same-gender interaction with an attractive examiner may be further supported by the *interpretation comparison model* (Stapel & Koomen, 2000; 2001). Based on empirical findings revealing that implicit automatic social comparisons led to contrast effects (Stapel & Suls, 2004), this model would also predict contrast effects for evaluative comparisons. After the self-estimation in the

beginning of the testing – which may be perceived as threatening for test takers when confronted with an attractive examiner – the assessment took its course and social comparison processes arose in interaction with someone comparable.

Summarizing, concerning the second and third research aims – i.e. how examiners were perceived with regard to gender stereotypes and investigation of preferences for examiners – a preference for female examiners can be derived. The majority of test takers preferred female examiners over male examiners. With reference to ascribe examiners' characteristics, the results of the third study are partly supportive as regards common stereotypes. In line with such stereotype related ascriptions, women as examiners were perceived as more social competent than men as examiners. Opposite to given stereotypes, the results contradict the male stereotype of more expertise and proficiency with revealing comparable ratings on expertise for middle-aged female and male examiners. Again and as referred to above, this may even be related to a feminine image of Psychology as a profession. With reference to expertise, age revealed to be a stronger social cue than gender in this study. Overall it can be concluded that stereotypes influence the perception of examiner's social competence and expertise. In turn, these stereotypical perceptions may serve as an explanation for examiner preferences.

Taking results out of Study 1 and Study 4 into account of interpretation of the given results, the preference for women as examiners is supported by better performance in presence of female examiners. Persons therefore tend to have a preference that is in line with higher performance or higher self-estimations as an indicator for metacognitions during testing. Not only ascribed higher levels of social competences, but also less threat and expected lower difficulty of task may serve as explanations and should be considered in further studies.

In the following, some limitations regarding these findings should be considered: First, the interpretation and informative value of the third study for practical purpose is limited, as no real testing took place after the questioning. Therefore, no conclusions concerning performance differences due to different examiner preferences can be drawn. Future studies should further investigate the impact of perception and evaluation of examiners on the actual performance of test takers.

Second, based on the results of the third study, the question remains if devaluation of young male examiners prevails a consistent observation. It is possible that the current debate concerning supposed discrimination of boys and young men in education (for an overview

and analysis see Hannover & Kessels, 2011) influenced the evaluations. Moreover, future studies have to investigate if this effect may be reduced by training young male examiners.

Third, in view of future research, it should be investigated whether socially competent examiners may lead to the expectation of a less threatening testing atmosphere, a more helpful and supporting examiner and overall, an easier testing atmosphere. Further research is needed to scrutinize these suggestions.

Referring to Study 2, it should be considered that the ratings of attractiveness were not derived by external observers, but referred to the individual ratings of the test takers. Such subjective perceptions were employed as they appeared more relevant to explain individual test takers' results. However, this advantage on the one hand may be accompanied by disadvantages on the other hand: For example, there is the possibility that individual test experiences had impact on attractiveness ratings. Furthermore we cannot exclude attractive examiners having behaved differently with reference to paraverbal or nonverbal behavior towards same-gender and opposite-gender test takers. Future studies should investigate possible differences between the perception of test takers and external observers.

Overall, this study may also stimulate research investigating the impact of perceived attractiveness in oral examinations.

### 6.1.3 Results Gained on Test Taker Characteristics Influencing how Examiners are Perceived

As already introduced, test taker characteristics may also contribute to the social interaction in face-to-face test settings. Results of a study by Danso and Esses (2001) concerning the influence of examiner's ethnicity on the performance of test takers with high SDO were reported in Chapter 1 and taken as a basis for Study 3 in this thesis. The third study is the first revealing that participants with high SDO tended to even stronger preferences for female examiners compared to participants with low SDO.

Higher preference for women as examiners of persons with a higher orientation towards thoughts of social dominance is well in line with the positive relationship shown between SDO and stereotyping and acceptance of traditional gender roles (Sidanius & Pratto, 2001; Tausch & Hewstone, 2010). Following this rationale it may be suggested that people possessing higher levels of SDO would prefer a female examiner because she would activate the own achievement motivation and be possibly perceived as easily to compete with. As assessment dyads always reflect status differences between examiner – with power and control over the situation – and the examinee – who is requested to provide personal

information or show a performance – future research has to consider SDO as a relevant person variable in testing.

Concerning the relevance of test taker's gender on interaction in face-to-face testing, prior research presented contradictory and inconclusive results (Rumenik et al., 1977). However, based on the current results, this thesis indicated no effect of test taker's gender on the preference or evaluation of examiners regarding social competence or expertise. This finding accords to social psychological research indicating no substantial gender differences in stereotypical ascriptions (e.g., Désert & Leyens, 2006).

However, test taker's gender had impact on the perception of examiner's attractiveness. The underlying processes were not investigated in this thesis. Moreover, male test takers with high SDO preferred female examiners even stronger than male test takers with low SDO, whereas the difference for female test takers with either high or low SDO was not so pronounced.

The given results also contradict a possible stereotype threat effect through examiner characteristics. Stereotype threat has been defined as affecting only those participants belonging to a stereotyped group concerning the specific performance domain (Steele & Aronson, 1995). The results of this thesis are not in line with results reported by Marx and Goff (2005) who found a stereotype threat effect for Black test takers when tested by a White examiner. Stereotype threat emerged in presence of a White examiner, whereas the presence of a competent Black examiner minimized the potential threat of the evaluative test. The studies of this thesis found no such performance handicap due to existing stereotypes elicited by the examiner. Different explanations could be proposed as regards this finding: On the one hand it could be proposed that test situations investigated in this thesis did not create stereotype threat effects. In this case, assessment mode or test setting may have not been containing gender related cues with reference to achievement. Nevertheless, in two studies a general knowledge test and in one study a computerized speed test were employed. In most domains on general knowledge men outscore women (Lynn, Wilberg, & Margraf-Stiksrud, 2004) and computerized tasks have been shown to reveal stereotype threat as well (Koch, Müller, & Sieverding, 2008; Smith, Morgan, & White, 2005). So considering that effects were too small again to be detected in the Study 1 and Study 2, also Study 4 did not show examinees' gender effects.

Subsuming, the fourth research aim concerned the contribution of test taker characteristics to ascribed examiners' characteristics, preferences for examiners, or achievement under male or female examiners in face-to-face test settings. Based on the results

of this thesis I assume that test taker's SDO influences the expectations of upcoming test situations and this person variable has to be further regarded in social interaction with status differences like face-to-face testing. The results of this thesis do not indicate a direct impact of test takers gender on examiner perception. Furthermore, the results suggest that other differences among test takers have to be considered as well. Individual characteristics as self concept and self esteem, cultural background, or test experience may influence the interaction of test taker and examiner in a specific testing. With regard to fairness future studies have to take individual differences between test takers into account more thoroughly.

## 6.2 Implications for the Testing Practice

Different implications for the testing practice may be derived altogether from the results of the single studies: As already introduced, in Psychological Assessment the use of standardized tests aims to meet certain quality standards. The findings of this thesis seem especially relevant for claims of objectivity and fairness.

### 6.2.1 Possible Conclusions on Examiner Effects on Objectivity and Fairness

As referred to in Chapter 1, objectivity in psychological testing is a key principle and precondition in Psychological Assessment. Usually, person characteristics of examiners have not been considered as influential in standardized testing.

However, this thesis adds new perspectives to the discussion of measurement quality by addressing examiners' possible influence on results gained in a testing even in a highly standardized setting: The finding that test takers estimate their verbal knowledge lower (Study 1) and perform worse on a speed test (Study 4) in presence of a male examiner compared to female examiner, the indication for performance decrements on a knowledge test when tested face-to-face by a male examiner (Study 1), and the outcome that perceived attractiveness of an examiner may lower the performance on a knowledge test (Study 2) supports the concern that objectivity in standardized test settings may be impaired. This thesis also highlights examiners' impact even if the test is presented in computerized form.

Effect sizes indicating higher cognitive performance in the presence of women were overall small. Nevertheless, this does not attenuate their relevance for assessment practice. For example, Willingham and Cole (1997) explicated that even small effect sizes may have large impact and very small group differences may produce great factual effects if only few persons are selected from a large population.

The effect size of results regarding examiners' gender on self-estimated verbal knowledge was almost medium. Persons estimated their knowledge, again, higher in presence of a female compared to a male examiner. This result does not only indicate that self-estimations may be even more impaired by presence of either a man or woman. Future studies have to show whether this result may also be extended to other forms of assessment, as for example, employment interviews where applicants are not requested to *show* their knowledge and competence but to *present and describe* it.

With reference to fairness, Chapter 1 addressed a broader concept of fairness by including group characteristics as well as psychological variables (Helms, 2006). Based on the given results, I do not conclude general fairness problems in this thesis with reference to the test taker characteristics including gender and age.

Concerning test takers' characteristics, only Study 3 revealed that test takers differing in their individual levels of SDO preferred female examiners. It could be assumed in line with Danso and Esses (2001) that a stronger endorsement of female examiners may also further lead to an improved performance in presence of women as examiners of test takers with high SDO. This would indicate a combined fairness - objectivity violation. However, this issue should also be investigated in future studies.

Once more it could be argued that the differences in outcomes were rather small in the presented studies. Concerning fairness it has to be kept in mind that these investigations only used mean differences. It may be proposed that additional subgroup characteristics may discover larger effects within subsamples. Future studies may, for example, reveal additional results on vulnerability regarding situational effects, e.g. based on test anxiety, or self esteem.

### 6.2.2 Implications for Test Use with Reference to Performance Testing

The relevance of the results for the test use depends on the context and aim of the specific testing. Objectivity and fairness are especially important in the context of educational or vocational aptitude tests. The following suggestions for test use concern primarily such test settings. I will discuss six suggestions derived from the results of this thesis.

First, in assessment settings where face-to-face testings are indispensable, it could be proposed that examiners' uniformity as regards gender may offer a solution to increase objectivity. Therefore the finding that men and women as examiners induce different chances for test takers would lead to the conclusion, that all test takers should be tested by either only men or only women in order to match the testing situation with reference to its objectivity. To address additional characteristics (as for example, attractiveness), even the same examiner for

all test takers could be suggested. The underlying assumption would be that examiner influences would then be the balanced out for all test takers. Indeed, the practicability of this suggestion seems questionable – especially if large testings are conducted. In line with this reasoning another approach refers to group testing: In situations where tests may be conducted in groups the presence of several examiners – female and male – may level potential examiner gender effects out. Furthermore, effects due to perceived attractiveness may be diminished. However as already discussed, further investigations concerning examiner effects in group testing are needed and after all, not all testing may be held in group settings.

A third suggestion, again in individual face-to-face testing, would refer to the option of choice: This strategy would be based on the finding that different test takers perceive examiners differently leading to different preferences concerning the examiner. One could assume in line with these considerations, that test takers intuitively know examiner characteristics benefiting their own performance. Under these assumptions, preferences would suggest the relevance of examinees' choice in testing procedures. Nevertheless, this solution bears disadvantages itself: First, it is not clear if test takers would in fact be capable of and interested in selecting their best-fitting examiner eliciting an optimum performance. Second, in real life testing it may not be possible to provide examiners for different preferences. After all, test takers may not only wish to choose the gender of the examiner, but also age, ethnicity, look, or any other characteristic. However, further studies may give inside into the relevance of further characteristics on performance.

Fourth, in cases where different examiners and face-to-face testings are indispensible necessities, standardization should at least be increased with reference to examiners behavior – in terms of verbal behavior as well as paraverbal and nonverbal behavior like intonation and gestures. Although all examiners in this study were individually trained – as in practice – no training of nonverbal behavior or speech training has been included. Nevertheless, the results of this thesis suggest that a focus on standardization of instructions and verbal interactions, question and answer mode is highly advisable.

The fifth proposition refers to test settings where an examining person could be disclaimed. Based on given results it cannot be concluded that computerized assessment excludes examiner effects. Future studies should show, whether they are at least reduced. However, also internet based testings may offer solutions in some contexts. Internet based assessment is becoming popular due to the wide distribution of the internet and the development of new test forms (e.g., *implicit association tests*, Greenwald, Nosek, & Banaji, 2003). Nevertheless, this possibility discloses other disadvantages as well because not all test

contents seem equally appropriate for unattended testing (e.g., Barak, 1999; Buchanan & Smith, 1999; Naglieri et al., 2004).

Overall, based on previous research, use of written tests may be superior compared to oral examinations (cf. Daelmans, Scherpbier, Van der Fleuten, & Donker, 2001; Pokorny & Frazier, 1966).

### 6.3 Concluding Remarks

With this thesis for the first time examiner gender effects were systematically investigated and were revealed for face-to-face settings using standardized contemporary psychological tests. For the first time, this result has been gained based on different studies that employed a considerable number of examiners minimizing the possible influence of individual examiners on test takers. Results of this thesis suggest effects of slightly higher performance as well as higher self estimated performance in the presence of a female examiner compared being tested by a male examiner.

The revealed examiner gender effects are especially noteworthy for Psychological Assessment in practice, especially as regards fields of test application where objectivity and fairness are most essential features. At least in Western societies it is to be expected that the need for Psychological Assessment in the domain of aptitude testing will rise (cf. Ortner, 2010). With regard to selection processes in companies or at school or universities Psychological Assessment may contribute to the establishment of adequate test procedures. This not only helps selecting appropriate applicants but also increases the chances of underrepresented groups to participate in education and powerful positions in society. For example, members of ethnical minorities are still underrepresented in higher education and women are rarely found in high-ranking positions in economy and politics (European Commission, 2011; Konsortium Bildungsberichterstattung, 2006). Therefore I propose examiner effects on test taker's performance being a serious problem for contemporary and future Psychological Assessment concerning educational, vocational, and training contexts.

However, in order to avoid examiner gender effects as revealed in this thesis in future, not only additional studies including a wider range of settings and measures are required. Future interventions would also benefit from a more deepened understanding of underlying interactional processes between test takers and the examiners in standardized testing.

Overall, the results of this thesis claim for a more careful consideration of the test situation itself. In future, research should further reveal which settings may allowing for

evaluation and interpretation of an individual's competencies – most undistorted by the conditions under which the individual was performing.

## 6.4 References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Barak, A. (1999). Psychological applications on the internet: A discipline on the threshold of a new millennium. *Applied & Preventive Psychology, 8*(4), 231-245. doi: 10.1016/S0962-1849(05)80038-1

Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology, 72*, 544-557. doi: 10.1037/0022-3514.72.3.544

Biernat, M., & Manis, M. (1994). Shifting Standards and Stereotype-Based Judgments. *Journal of Personality and Social Psychology, 66*, 5-20. doi: 10.1037/0022-3514.66.1.5

Bookout, D. V., & Hosford, R. E. (1969). Administration effects on the S-329 of the GATB using three experimental treatments. *Journal of Employment Counseling, 6*, 124-133.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology, 90*, 125-144. doi: 10.1348/000712699161189

Campbell, J. D., & Tesser, A. (1985). Self-evaluation maintenance processes in relationships. In Duck, S., & Perlman, D. (eds.) *Understanding Personal Relationships: An Interdisciplinary Approach*. Sage: London

Cohen, J. (1992). A Power Primer. *Psychological Bulletin, 112*, 155-159.

Daelmans, H. E. M., Scherpbier, A., Van der Vleuten, C. P. M., & Donker, A. J. M. (2001). Reliability of clinical oral examinations re-examined. *Medical Teacher, 23*, 422-424. doi: 10.1080/01421590120042973

Danso, H. A., & Esses, V. M. (2001). Black experimenters and the intellectual test performance of White participants: The tables are turned. *Journal of Experimental Social Psychology, 37*, 158-165. doi: 10.1006/jesp.2000.1444

Désert, M., & Leyens, J.-P. (2006). Social comparisons across cultures: I. Gender stereotypes in high and low power distance cultures. In S. Guimond (Ed.), *Social Comparison and social psychology: Understanding cognition, intergroup relations, and culture* (pp. 303-317). Cambridge, England: University Press.

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review, 1,* 3-14. doi: 10.1016/j.edurev.2005.11.001

Efklides, A., Samara, A., & Petropoulou, M. (1999). Feeling of difficulty: An aspect of monitoring that influences control. *European Journal of Psychology of Education, 14*, 461-476.

European Commission (Ed.). (2011). *Report on Progress on Equality between Women and Men in 2010 - The gender balance in business leadership*. Luxembourg: Publications Office of the European Union, 2011. doi: 10.2767/99441

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses *Behavior Research Methods, 41*, 1149-1160.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117–140. doi: 10.1177/001872675400700202

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197-216. doi: 10.1037/0022-3514.85.2.197

Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin, 131*(6), 898-924. doi. 10.1037/0033-2909.131.6.898

Hannover, B., & Kessels, U. (2011). Sind Jungen die neuen Bildungsverlierer? Empirische Evidenz für Geschlechterdisparitäten zuungunsten von Jungen und Erklärungsansätze. *Zeitschrift für Pädagogische Psychologie, 25*(2), 89-103. doi: 10.1024/1010-0652/a000039

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial group assessment: A quantitative perspective. *American Psychologist, 61*(8), 845-859. doi: 10.1037/0003-066X.61.8.845

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science, 11*, 365-371. doi: 10.1111/1467-9280.00272

Karremans, J. C., Verwijmeren, T., Pronk, T. M., & Reitsma, M. (2009). Interacting with women can impair men's cognitive functioning. *Journal of Experimental Social Psychology, 45*, 1041-1044. doi: 10.1016/j.jesp.2009.05.004

Koch, S. C., Müller, S. M., & Sieverding, M. (2008). Women and computers. Effects of stereotype threat on attribution of failure. *Computers & Education, 51*, 1795-1803. doi: 10.1016/j.compedu.2008.05.007

Konsortium Bildungsberichterstattung (Eds.) (2006). *Bildung in Deutschland. Ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration*. Bielefeld: Bertelsmann Verlag. Retrieved 9.7.2011 from http://www.bildungsbericht.de.

LaFrance, M., Hecht, M. A., & Paluck, E. L. (2003). The contingent smile: A meta-analysis of sex differences in smiling. *Psychological Bulletin, 129*, 305-334. doi: 10.1037/0033-2909.129.2.305

Lüdtke, O., Robitzsch, A., Trautwein, U., Kreuter, F., & Ihme, J. M. (2007). Are there test administrator effects in large-scale educational assessments: Using cross-classified multilevel analysis to probe for effects on mathematics achievement and sample attrition. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*, 149-159. doi: 10.1027/1614-2241.3.4.149

Lynn, R., Wilberg, S., & Margraf-Stiksrud, J. (2004). Sex differences in general knowledge in German high school students. *Personality and Individual Differences, 37*, 1643-1650. doi: 10.1016/j.paid.2004.02.018

Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology, 44*, 645-657. doi: 10.1348/014466604X17948

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., et al. (2004). Psychological Testing on the Internet: New Problems, Old Issues. *American Psychologist, 59*, 150-162. doi: 10.1037/0003-066X.59.3.150

Olos, L., & Hoff, E.-H. (2006). Gender ratios in European psychology. *European Psychologist, 11*, 1-11. doi: 10.1027/1016-9040.11.1.1

Ortner, T. M. (2010). Das Potenzial Psychologischer Diagnostik im Angesicht aktueller gesellschaftlicher Herausforderungen. In: Kubinger, K.D. & Ortner, T.M. (Eds.) *Psychologische Diagnostik in Fallbeispielen.* Göttingen: Hogrefe.

Socio-economic Panel (2010). (SOEP) data for the years1984-2009, version 26.

Pokorny, A. D., & Frazier, S. H. (1966). An Evaluation of Oral Examinations. *Journal of Medical Education, 41*, 28-40.

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research* (enlarged ed.). New York, N.Y.: Irvington Publishers, Inc.

Rumenik, D. K., Capasso, D. R., & Hendrick, C. (1977). Experimenter sex effects in behavioral research. *Psychological Bulletin, 84*, 852-877. doi: 10.1037/0033-2909.84.5.852

Samuel, W. (1977). Observed Iq as a Function of Test Atmosphere, Tester Expectation, and Race of Tester - Replication for Female Subjects. *Journal of Educational Psychology, 69*, 593-604. doi: 10.1037/0022-0663.69.5.593

Sattler, J. M., & Theye, F. (1967). Procedural, situational, and interpersonal variables in individual intelligence testing. *Psychological Bulletin, 68*, 347-360. doi: 10.1037/h0025153

Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology, 39*, 68-74. doi: 10.1016/S0022-1031(02)00508-5

Sidanius, J., & Pratto, F. (2001). *Social Dominance* Cambridge: University Press.

Smith, J. L., Morgan, C. L., & White, P. H. (2005). Investigating a Measure of Computer Technology Domain Identification: A Tool for Understanding Gender Differences and Stereotypes. *Educational and Psychological Measurement, 65*, 336-355. doi: 10.1177/0013164404272486

Spence, J. T., & Buckner, C. E. (2000). Instrumental and expressive traits, trait stereotypes, and sexist attitudes. *Psychology of Women Quarterly, 24*, 44-62. doi: 10.1111/j.1471-6402.2000.tb01021.x

Stapel, D. A., & Koomen, W. (2000). Distinctness of others and malleability of selves: Their impact on social comparison effects. *Journal of Personality and Social Psychology, 79,* 1068–1087. doi: 10.1037/0022-3514.79.6.1068

Stapel, D. A., & Koomen, W. (2001). I, we, and the effects of others on me: How self-construal moderates social comparison effects. *Journal of Personality and Social Psychology, 80,* 766–781. doi: 10.1037/0022-3514.80.5.766

Stapel, D. A., & Suls, J. (2004). Method matters: Effects of implicit versus explicit social comparisons on activation, behavior, and self-views. *Journal of Personality and Social Psychology, 87*, 860–875. doi: 10.1037/0022-3514.87.6.860

Steckler, N. A., & Rosenthal, R. (1985). Sex-Differences in Nonverbal and Verbal Communication with Bosses, Peers, and Subordinates. *Journal of Applied Psychology, 70*, 157-163. doi: 10.1037/0021-9010.70.1.157

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797-811. doi: 10.1037/0022-3514.69.5.797

Tausch, N., & Hewstone, M. (2010). Social Dominance Orientation Attenuates Stereotype Change in the Face of Disconfirming Information. *Social Psychology, 41*, 169-176. doi: 10.1027/1864-9335/a000024

Thornton, B., & Moore, S. (1993). Physical Attractiveness Contrast Effect - Implications for Self-Esteem and Evaluations of the Social Self. *Personality and Social Psychology Bulletin, 19*, 474-480. doi: 10.1023/A:1018867409265

Willingham, W. W., & Cole, N. S. (1997). Introduction. In W. W. Willingham & N. S. Cole (Eds.), *Gender and Fair Assessment* (pp. 1-15). Mahwah, NJ: Erlbaum.

# Appendix

## List of Tables



## List of Figures

# Curriculum Vitae

For reasons of data protection, the curriculum vitae is not included in this version.

# List of Publications

Ortner, T. M. & Vormittag, I. (2011). Test administrator's gender affects female and male students' self-estimated verbal general knowledge. *Learning and Instruction, 21,* 14-21. doi: 10.1016/j.learninstruc.2009.09.003

Vormittag, I. & Ortner, T. M. (submitted). Too Perfect to Challenge: Effects of Attractive Examiners on Performance of Men and Women.

Vormittag, I. & Ortner, T. M. (submitted). Does Gender Speak Louder than Words? How Stereotypes Influence Perceptions of and Preferences for Test Examiners.

Vormittag, I. & Ortner, T. M. (submitted). Better Cognitive Functioning in the Presence of Women! Situational Influence on Test Performance in Large-Scale Assessment.

# Erklärung zur Dissertation

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbständig verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht verwendet. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, Juli 2011 Isabella Vormittag