Making the invisible visible : essays on overconfidence, discrimination and peer effects

Making the Invisible Visible

Essays on Overconfidence, Discrimination and Peer Effects

Published by ROA
Postbus 616
6200 MD Maastricht

ISBN: 978-90-5321-530-2

Printed in the Netherlands by Océ Business Services

Making the Invisible Visible

Essays on Overconfidence, Discrimination and Peer Effects

Dissertation

to obtain the degree of Doctor at
Maastricht University,
on the authority of Rector Magnificus,
Prof. dr. L.L.G. Soete
in accordance with the decision of the Board of Deans,
to be defended in public on
Friday October 24th, 2014 at 12:00 hours

by

Jan Feld

Supervisor

Prof. dr. A. de Grip

Co-supervisor

Prof. dr. T. Dohmen, University of Bonn

Assessment Committee

Prof. dr. R.K.W. van der Velden (chair)
Dr. A.S. Booij (University of Amsterdam)
Dr. O. Marie
Prof. dr. D. Webbink (Erasmus University Rotterdam)

Acknowledgements

I felt very much at home in Maastricht and the years of my PhD have so far been the best of my life. I enjoyed the intellectual discussions, the great working climate at ROA and the welcoming and inspiring Maastricht University PhD community. I would like to thank the many people who made this thesis possible and my time here so very enjoyable.

Andries, you have been a great supervisor. You are very open, helpful, supportive and incredibly calm and kind. Maria, Ahmed and Nico: you are the best office mates one could hope for. I enjoyed our discussion as much as the bullshitting. Maria, you have been a great office and house mate. Ahmed, I really enjoyed our discussion about religion. Thank you for being so good to your "haram" friend. Nico, you have been a great friend and coauthor. I loved our discussions that often started with "Nico, I have been thinking…" during lunch, coffee breaks or over a beer. I admire your enthusiasm and knowledge. It was in our discussions that I learned most about economics! Ulf, it has been great working with you. Your hard work and determination really pushed our project forward. I also enjoyed that we managed to find the time to take emergency breaks when necessary. Dan, Nico and I agree that walking into your office was the best idea of our PhD time. It is so much fun working with you and we have learned a lot from you. Thank you for inviting me to spend some time at the University of Texas at Austin.

I would also like to thank the people at Maasticht University who provided me with data and valuable background information: Joël Castermans, Jeannette Hommes, Ad van Iterson, Paul Jacobs, Christian Kerckhoffs, Sylvie Kersten, Sanne Klasen, Caroline Kortbeek, Kim Schippers and Alexander Vostroknutov. I really appreciate your effort, patience and good will. Without you this thesis would not have been possible.

Outside of work I would like to thank my friends who made my time in Maastricht so special. Melissa and Simone: you are awesome! Thank you for being such good friends. I would also like to thank my improv theatre friends who made Thursdays my favorite day of the week: Jessie, Mehrdad, Nevena, Gabri, Anna, Sanne, Paul, Nadine, Nordin, Elsa, Jeroen, Gintare, Shuan, Alejandro, Julie, Johannes, Popi, Despina, Jasper, etc. – you rock!

Last but not least I would like to thank my family. Marita, Andreas and Timo: thank you for always being there for me. Knowing that I can always count on you is the foundation from which I go out and try to conquer the world.

Content

List of Figures

List of Tables

Chapter 1
Introduction to Thesis

1.1 Introduction

1.1.1 Motivation

As social scientists our goal is to better understand the social world by uncovering laws that govern human behavior. Uncovering these laws, however, is often challenging as we live in a complex social world where many phenomena are interconnected, and not always in a straightforward way. This makes it difficult to distinguish correlation from causation. Furthermore, we are often interested in studying abstract concepts which cannot be directly observed at all. In other words, we face empirical challenges to answer our research questions. As empirical social scientists we therefore often have to get creative by finding, collecting and analyzing data to overcome these obstacles. Our job is thus to cut through the complex social world and uncover its hidden relationships – to use creative identification strategies to make the invisible visible.

1.1.2 Aim

In this thesis, I uncover relationships in the domains of overconfidence, discrimination, and peer effects in education. The common denominator of all chapters is that they use sophisticated identification strategies. Through these identification strategies, I will answer three different questions: What is the relationship between skill and overconfidence? Are differences in outcomes driven by discrimination or favoritism? How does peer ability affect student achievement?

1.1.3 Overconfidence

The Dunnig-Krueger effect (DK effect) proposes that skill and overconfidence are generally inversely related, with the low skilled being most overconfident while the high skilled being accurate on their skill assessments (Kruger and Dunning 1999). Estimating the DK effect, however, is difficult because both skill and overconfidence are unobservable. Researchers instead use (test) performance and overestimation, which is the difference between expected and realized performance, as their respective measures. Many studies have shown that, in different populations and tasks, low performers usually vastly overestimate their performance while high performers have a more accurate view on their performance (Kruger and Dunning 1999, Ehrlinger, Johnson et al. 2008, Ryvkin, Krajč et al. 2012, Schlösser, Dunning et al. 2013). However, performance measures skill with some error which can be

considered as luck in a test. This error, which is part of the performance measure, is often at the same time part of the overestimation measure. This fact alone can create a negative relationship between performance and overestimation, even if there is no systematic relationship between skill and overconfidence (Krueger and Mueller 2002). Intuitively, think of a student with bad luck on a test: This bad luck will reduce her performance and, at the same time, increase her overestimation, thus making her appear as less skilled and more overconfident. In this thesis, I show that the true relationship between skill and overconfidence can be estimated with an instrumental variable approach by using a second performance measure as an instrument for a person's first performance measure.

1.1.4 Discrimination

Economists have been interested in studying discrimination at least since Becker's (1957) theory of discrimination. But are differences in outcomes caused by people discriminating against others – exophobia – or are they a result of people favoring their own kind – endophilia? Distinguishing between exophobia (out-group discrimination) and endophilia (in-group favoritism) is difficult because differences in outcomes could be driven by either, or by both. We show that one can disentangle exophobia and endophilia by comparing the treatment of randomly selected members the same group (e.g. females) with and without visible characteristics which reveal to which group they belong. We do this in a field experiment in which we randomly show and conceal students' names on exams. The students without visible names on their exams then serve as a neutral baseline against which we measure exophobia and endophilia. We then have evidence for exophobia when students who belong to a different group as the grader (e.g. have a different gender) receive lower grades when their names are visible. Conversely, we have evidence for endophilia when students who belong to the same group than the grader (e.g. have the same gender) receive higher grades when their names are visible.

1.1.5 Peer Effects in Education

There is a large literature in economics which studies how peer ability affects student performance (Hoxby 2000, Angrist and Lang 2004, Carrell, Fullerton et al. 2009, Duflo, Dupas et al. 2011, Burke and Sass 2013, Carrell, Sacerdote et al. 2013). However, identifying peer effects in education is empirically challenging because students are usually in a peer group (school, grade, tutorial group, dorm, etc.) for a specific reason (e.g. their ability, past performance, motivation) and it is difficult for the researcher to distinguish between the reason for belonging to a peer group and

the actual peer effect. Therefore, this selection problem can be solved by randomly assigning students to peer groups, like it is done at the School of Business and Economics (SBE) of Maastricht University. We exploit this random assignment at the SBE to estimate the effect of being assigned to peers with different abilities.

1.2 Outline

In Chapter 2 we discuss how to estimate the relationship between skill and overconfidence. The DK effect states that the low-skilled are generally overconfident while the high-skilled are more accurate (Kruger and Dunning 1999). This effect seems to be widely accepted in the scientific literature, to a large extent because many studies have shown that bottom performers usually vastly overestimate their own performance, while top performers are more accurate (Kruger and Dunning 1999, Ehrlinger, Johnson et al. 2008, Ryvkin, Krajč et al. 2012, Schlösser, Dunning et al. 2013). However, both performance and overestimation measure skill and overconfidence with some error. This measurement error may cause the DK effect to be a statistical artifact. In this chapter, we discuss potential biases when estimating the DK effect, the shortcomings of the currently used bias correction methods, and propose an alternative way to estimate the DK effect by using the performance on an independent task as an instrument for a person's first performance measure. We show that this instrumental variable approach allows for a consistent estimate of the true relationship between skill and overconfidence.

In Chapter 3 we empirically estimate the DK effect using the instrumental variable method discussed in Chapter 2 using exam grades and grade predictions data from economics students. In this context, we use a natural second performance measure, the student's past grade point average, as an instrument. Our results are consistent with the DK effect hypothesis: On average, low skilled students are more overconfident while high skilled students are more accurate. We further show that the size of these instrumental variable estimates is substantially different to the size of estimates with the estimation methods used in the current literature.

In Chapter 4 we investigate whether differences in exam grades are driven by discrimination or favoritism. The discrimination literature usually estimates relative outcomes. But does a differential arise because agents discriminate against others – exophobia - or because they favor their own kind - endophilia? We answer this question in the context of a field experiment at the SBE, which has large shares of

Dutch and German students. In this context, we assigned graders randomly to students' exams that did or did not contain names. We were thus able to infer whether the student they were grading matched or not his/her nationality or gender. We find on average endophilia but no exophobia by nationality, but neither by gender. We identify distributions of graders' endophilic and exophobic preferences. Furthermore, we show that a changing correlation between endophilic and exophobic preferences generates changes in market (wage) differentials.

In Chapter 5 we investigate the effect of peers' ability on university students' achievement. We use data on economics and business students who are randomly assigned to tutorial groups at the SBE. Being assigned to tutorial groups peers with higher ability, as measured by students' past GPA, leads to very small increases in student grades in the linear-in-means specification. These results, however, hide some heterogeneity: low ability students perform worse when exposed to a larger share of high ability students. Our findings point to an inverse U-shaped relationship between performance and peer ability: students benefit from better performing peers as long as the difference between their own and peer ability does not exceed a certain threshold.

In Chapter 6 I summarize the main findings of the thesis. Chapter 7 discusses the valorization in terms of policy implications of the thesis.

Chapter 2
Unskilled and Unaware? On Estimating the Relationship between Skill and Overconfidence*

* This chapter is based on joined work with Jan Sauermann and Andries de Grip. We thank Thomas Dohmen, Jonas Lang, Nicolas Salamanca, and the participants of the DuHR seminar for valuable comments.

2.1 Introduction

The Dunning-Kruger effect (DK effect) states that the low-skilled are on average overconfident while those who are high-skilled are more accurate in assessing their skill. The DK effect has received a lot of attention in the scientific literature: the seminal article by Kruger and Dunning (1999) has been cited more than 1,900 times.[1] Apart from the psychological literature many researchers in other scientific disciplines seem to have accepted the DK effect as a psychological fact that can be used to explain individuals' behavior (for example in law (Tor, 2002), management science (Dane & Pratt, 2007), medicine (Haun, Zeringue, Leach, & Foley, 2000)). Kruger and Dunning (1999) argued that the lack of metacognitive skills of the unskilled can explain the DK effect. The intuition behind this explanation is that the skills that are required to perform well are often the same skills that are required to evaluate one's skill accurately. And when not being able to evaluate one's skill accurately people tend to be overconfident.

In this chapter, we show that a closer look at the statistical methods currently used to demonstrate this apparent fact does not allow for this conclusion. The main reason why the DK effect has nevertheless been so widely accepted is probably that many studies have shown a pattern of high overestimation of low performing individuals, while high performers are more accurate (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999; Kruger & Dunning, 2002). However, performance and overestimation measure skill and overconfidence with error. Krueger and Mueller (2002) have argued that this measurement error causes *regression effects* which combined with the better-than-average heuristic, can create the observed pattern between performance and overestimation even when skill and overconfidence are unrelated.[2] In this chapter, we show that even the currently used correction methods do not take estimation biases fully into account.

The DK effect has been studied with absolute measures (e.g. test points) as well as relative measures (e.g. percentiles) of performance. We will show that the choice between absolute and relative measures has important implications for estimating the DK effect. When using absolute measures, there are two potential biases: *regression effects* and *attenuation bias*. The reason for regression effects is that performance measures skill with error. This error, which can be interpreted as (bad)

1. As we write this on 29.07.2014 Google Scholar shows that the article has been cited 1,975 times.
2. This explanation has later been generalized by Burson, Larick et al. (2006) in their noise-plus-bias model which states that over- and underestimation are a result of random error, which causes regression effects, and a task-induced bias, which shifts the overall level of overconfidence.

luck, is part of the dependent variable (overestimation) as well as the independent variable (performance) and artificially creates a negative relationship between performance and overestimation. To illustrate this, consider a student with bad luck on a test: bad luck will reduce her performance and increase her overestimation. Hence, the student appears to be less skilled and more overconfident. A second potential bias, which has received less attention in the literature, is attenuation bias: If performance measures skill with random error the absolute magnitude of the true relationship between skill and overconfidence will be underestimated. This means that if there is a negative relationship between skill and overconfidence, as proposed by the DK effect, attenuation bias will lead to an underestimation of the true relationship. When using relative measures of performance like percentiles, the (implicit) transformation into relative measures creates a negative correlation between skill and the error with which performance measures skill. We will show that this negative correlation further complicates the analysis.

We contribute to the literature on testing the DK effect in two ways. First, we carefully document the biases associated with estimating the relationship between skill and overconfidence. The aim of this is to provide a framework to discuss and interpret the results of previous studies, to understand their shortcomings, and to see in which direction the estimates are likely to be biased. Second, we introduce an alternative estimation method that requires the same data and less restrictive assumptions as previously used correction methods: instrumental variable (IV) estimation. We will show under which assumptions the IV method will lead to a consistent estimate of the relationship between skill and overconfidence.

The remainder of the chapter is structured as follows: In Section 2.2, we provide the framework in which the DK effect is being tested, the key variables, and the potential biases when testing the DK effect with absolute and relative measures of performance. Section 2.3 discusses the shortcomings of the currently used bias correction methods. In Section 2.4, we introduce the IV method and show under which conditions we will get a consistent estimate of the relationship between skill and overconfidence. Section 2.5 concludes.

2.2 The Framework, Key Variables and Biases

2.2.1 The Framework

The basic setup of DK effect studies is that individuals are asked to participate in a test, and are asked to guess their performance on this test. Expected performance is elicited either before or after the test and either in absolute terms (e.g. grades) or relative to their peers (e.g. percentiles). In the early studies, researchers have then shown the mean overestimation by different performance quartiles (Kruger & Dunning, 1999). A general finding is that the bottom quartile performers on average vastly overestimate their performance while the top quartile performers on average slightly underestimate their performance (Ehrlinger et al., 2008; Kruger & Dunning, 1999; Ryvkin, Krajč, & Ortmann, 2012; Schlösser, Dunning, Johnson, & Kruger, 2013). Krueger and Dunning (1999) explain this pattern by differences in metacognitive skills between the low and high-skilled. The intuition behind this explanation is that the skills necessary to perform well are often the same skills that are required to evaluate one's performance accurately. Therefore the low-skilled are overconfident while the high-skilled are more accurate about their absolute skill. However, due to the false consensus effect (Ross, Greene, & House, 1977), the high-skilled overestimate the skill of the others, and are therefore underconfident in their relative skill.

We model overconfidence as follows:

$$oc=\alpha+ \beta_s s+u_{oc} \qquad (1)$$

Where overconfidence (oc) is the sum of a constant term, α, and a variable component, $\beta_s s$, that depends linearly on the individual's skill level (s); u_{oc} is an idiosyncratic error term which captures individual differences in overconfidence that are unrelated to skill. We assume that skill is strictly exogenous, i.e. that $E[u|s_{oc}] = 0$. If both overconfidence and skill would be observed, ordinary least squares (OLS) regression of overconfidence on skill would lead to unbiased estimates of α and β_s.[3]

Looking at estimates of α and β_s jointly provides a simple framework to test the DK effect. The DK effect predicts that overconfidence declines with skill, i.e. that β_s is negative. The DK effect further predicts that the self-assessment errors are

3. Throughout the chapter, we assume that all variables are uncorrelated with u_{oc}, i.e. there is no other unobserved variable which is correlated with the explanatory variable and the dependent variable.

asymmetric, i.e. overconfidence of the low-skilled ($\alpha + \beta_s * s$ for low values of s) is large and positive while overconfidence of the high-skilled ($\alpha + \beta_s * s$ for high s) is small in absolute size. In the case of relative measures, this effect is negative for individuals with high values of s.

As the focus of this chapter is on estimating β_s, we will use $\frac{Cov\,(s,oc)}{Var(s)}$, which is equal to the skill coefficient of the hypothetical regression of overconfidence on skill, as a benchmark against which we evaluate the different estimation methods.[4] We discuss potential biases of estimates of α in Section 2.4.

2.2.2 The Key Variables

The four key variables of studies analyzing the DK effect are skill, performance, overconfidence and overestimation. We will here look at these variables on absolute scales. As we will discuss in Section 2.2.4, using relative measures complicates the analysis.

The understanding of skill in the DK effect literature is domain specific. The DK effect has been studied in a wide variety of domains ranging from understanding humor to gun safety knowledge (Ehrlinger et al., 2008; Kruger & Dunning, 1999). We therefore define skill as domain specific ability.

Performance, however, does not only reflect skill but also on luck. In this context luck captures all factors other than skill which influence test performance including environmental conditions (questions on the test, room temperature, noise, etc.) and physiological conditions (oxygen saturation in the brain, hormone levels, etc.). We thus model performance (p) as the sum of skill (s) and a random error component (ε_p) which is assumed to have zero mean and to be independent of the individual's overconfidence and skill:

$$p = s + \varepsilon_p \qquad (2)$$

For testing the DK effect, it is important to distinguish between overconfidence and overestimation. We define overconfidence as the difference between self-assessed and actual skill and understand overconfidence as a psychological bias. Overconfidence, however, is unobservable and we can instead observe overestimation which is the difference between expected and actual performance.

4. Recall that the OLS estimator of an regression of y on x can be written as $\hat{\beta} = \frac{Cov\,(x,y)}{Var(x)}$.

The important distinction between these two concepts is that overestimation, because it is calculated using the performance measure, is also driven by luck. To see that this matters, consider the following example: A person is asked to estimate how often out of ten times she can predict whether a fair coin lands on heads or tails. She has an accurate assessment of her skill and states that she can predict five out of ten tosses correctly. When predicting the coin tosses she, due to bad luck, only predicts four out of ten tosses correctly. This overestimation, however, is a result of bad luck and does not reflect a psychological bias. In this case her overestimation is one and her overconfidence is zero. This definition of overconfidence best reflects the understanding of the DK effect as a psychological and not merely statistical phenomenon.

We assume that expected performance (p_{exp}) is the person's best guess of her actual performance in a specific test and define expected performance as the sum of actual skill and overconfidence:[5]

$$p_{exp} = s + oc \tag{3}$$

When decomposing overestimation (oe) into its respective elements one can see that it is equal to the difference of oc and ε_p:

$$\begin{aligned} oe &= p_{exp} - p \\ oe &= (s + oc) - (s + \varepsilon_p) \\ oe &= oc - \varepsilon_p \end{aligned} \tag{4}$$

This means that overestimation is equal to individual's overconfidence about their skill level minus their luck on the test.

5. This assumption is more realistic when expected performance is elicited before the performance test. When expected performance is elicited after the performance test, the individual might already have some idea about her test performance (ε_p) and incorporate this in her expectation. We will discuss how such an adjustment affects the IV estimates in Section 2.4.2.

2.2.3 Biases When Estimating β_s with Absolute Measures

To show that the estimate of β_s from an OLS regression with overestimation and performance as measures for overconfidence and skill will be biased[6], we substitute overconfidence by overestimation and skill by performance in Equation (1):

$$\begin{aligned} oe &= \alpha + \beta_s p + u_{oc} \\ (oc - \varepsilon_p) &= \alpha + \beta_s (s + \varepsilon_p) + u_{oc} \end{aligned} \tag{5}$$

One can see from Equation (5) that the error component ε_p causes the two biases.[7] The first bias arises because the ε_p enters both the left-hand side variable (oe) and the right-hand side variable (p). This will lead to a negative bias of β_s, i.e. it causes the estimated coefficient to be more negative than the true β_s. One can illustrate this by seeing what happens for negative values of ε_p: Bad luck on a test will decrease performance and increase overestimation and will hence make unlucky individuals appear as less skilled and more overconfident. In accordance with the previous literature we call this bias *regression effects* (Burson, Larrick, & Klayman, 2006; Krueger & Mueller, 2002; Kruger & Dunning, 1999).

The second bias arises because performance measures skill with random error. This bias is referred to as *attenuation bias* (Hausman, 2001) and biases the estimated coefficient towards zero. If the true β_s is negative, as suggested by the DK effect, this will lead to a positive bias, i.e. it causes the estimated coefficient to be less negative than the true coefficient.

Another way to illustrate the direction of the two biases is to rewrite the OLS-estimator $\hat{\beta}_{OLS}$ as follows:

$$\hat{\beta}_{OLS} = \frac{Cov(p, oe)}{Var(p)} = \frac{Cov\,(s + \varepsilon_p, oc - \varepsilon_p)}{Var(s + \varepsilon_p)} \tag{6}$$

$$\hat{\beta}_{OLS} = \frac{Cov(s, oc)}{Var(s) + Var(\varepsilon_p)} - \frac{Var(\varepsilon_p)}{Var(s) + Var(\varepsilon_p)}$$

6. Testing the DK effect by showing average overestimation by performance quartiles, as done by Kruger and Dunning (1999), suffers in principle from the same biases as estimating it with OLS regression.
7. As mentioned in Footnote 3, we assume that there is no other unobserved variable which is correlated with the explanatory variable and the dependent variable.

Equation (6) shows that the estimated $\hat{\beta}_{OLS}$ is biased in two ways. First, the term $Var(\varepsilon_p)$ in the denominator of the first term reduces the absolute value of the $\hat{\beta}_{OLS}$ and thus biases the estimator towards zero. This is the attenuation bias and its magnitude depends on $Var(\varepsilon_p)$, i.e. how well the test measures skill, and it is proportional to the true β_s. Second, the second term shows the regression effects.[8] It is always negative and its magnitude depends on $Var(\varepsilon_p)$ but is independent of β_s. The size of the effect of the two biases together is *a priori* not clear.[9]

2.2.4 Biases When Estimating β_s with Relative Measures

Most DK effect studies use relative measures such as percentiles (Burson et al., 2006; Ehrlinger et al., 2008; Kruger & Dunning, 1999; Kruger & Dunning, 2002; Ryvkin et al., 2012). The problem with relative measures, which we denote with subscript r, is that the relative error (ε_{rp}), i.e. relative performance (p_r) minus relative skill (s_r), is bounded at the bottom and top of the skill distribution. Consider, for example, individuals at the bottom of the (relative) skill distribution. Their performance can only be the same or higher than their skill; their (relative) error can thus only be positive. Individuals at the top of the skill distribution, in contrast, can only have negative errors. This means that even under the standard assumption that the absolute performance measures skill with random error, the use of relative performance measures will artificially create a negative correlation between skill and the error term, i.e. $Cov(s_r, \varepsilon_{rp}) < 0$. This negative correlation will lead to a more complex bias of the relative OLS estimator $\hat{\beta}_{rOLS}$:

$$\hat{\beta}_{rOLS} = \frac{Cov(p_r, oe_r)}{Var(p_r)} = \frac{Cov\,(s_r + \varepsilon_{rp}, oc_{rp} - \varepsilon_{rp})}{Var(s_r + \varepsilon_{rp})} \tag{7}$$

$$\hat{\beta}_{rOLS} = \frac{Cov(s_r, oc_r) - Cov(s_r, \varepsilon_{rp}) + Cov(\varepsilon_{rp},\ oc_{rp}) - Var(\varepsilon_{rp})}{Var(p_r)}$$

8. The term $-\frac{\mathrm{Cov}(\mathrm{s},\varepsilon_\mathrm{p})}{\mathrm{Var}(\mathrm{p})}$ is dropped because with absolute measures we assume that skill and the error are uncorrelated. The term $\frac{Cov(\varepsilon_p, oc)}{Var(p)}$ is dropped because we assume that the error is uncorrelated with overconfidence.
9. Yet another way of showing the biases is to see how the explanatory variable, performance, is correlated with the error term. The error term of Equation (5) is equal to $u_{oc} + \varepsilon_p - \beta_s\varepsilon_p$. The first term is the idiosyncratic component. The estimate is biased because performance is correlated with the second component, which causes regression effects, and with the third component which causes attenuation bias.

$$\hat{\beta}_{rOLS} = \frac{Cov(s_r, oc_r)}{Var(p_r)} - \frac{Var(\varepsilon_{rp})}{Var(p_r)} - \frac{Cov(s_r, \varepsilon_{rp})}{Var(p_r)} + \frac{Cov(\varepsilon_{rp},\ oc_{rp})}{Var(p_r)}$$

Equation (7) shows that there are two additional terms when using relative measures. The first additional term is $-\frac{Cov(s_r,\varepsilon_{rp})}{Var(p_r)}$ which is due to the artificially negative correlation between relative skill and the error. The second additional term is $\frac{Cov(\varepsilon_{rp},oc_{rp})}{Var(p_r)}$. If there is a correlation between skill and overconfidence, the relative performance error is correlated with overconfidence through its correlation with skill. Consider, for example, people with a positive relative performance error. These people have a disproportionately low skill level – because of the negative errors are bounded – and – if the DK effect is correct – will also be more overconfident. Therefore the performance error will also be correlated with overconfidence if $\beta_s \neq 0$.

Another difference to the case of absolute measures in Equation (6) is that the denominator $Var(p_r)$ is equal to $Var(s_r)$. When skill is measured with random error, this error variance will increase the variance of the absolute performance measure. Transforming this measure from absolute into relative – by making a ranking like percentiles - compresses these differences. Consider, for example, three people who take a test: the ranks will be one, two and three in terms of both unobserved skill and observed performance. Hence, the variance of the two rankings will also be the same. As relative performance does not have a larger variance than relative skill, all terms in Equation (7) will be less attenuated than in Equation (6).

As $Var(p_r)$ is equal to $Var(s_r)$, the first term is equal to the unbiased estimator. The second term is biased due to regression effects, which leads to a negative bias. The third term leads to a positive bias because $Cov(s_r, \varepsilon_{rp})$ is negative as $-Cov(s_r, \varepsilon_{rp})$ is positive. The sign of the fourth term depends on $Cov(s_r, oc_r)$: when $Cov(s_r, oc_r) < 0$, as predicted by the DK effect, then $Cov(\varepsilon_{rp}, oc_{rp})$ is positive. As with absolute measures, the direction of the overall bias is *a priori* unclear. However, with these two additional terms, the properties of the bias become considerably more complex. This means that if absolute data are available, the use of absolute measures should be preferred as the biases will then be less complex, and can, as we will see in Section 2.4, be corrected for when using an appropriate method.

2.3 Currently Used Bias Corrections

2.3.1 The Split Sample Method

Regression effects are a result of the fact that it is the same error component which enters the performance and overestimation measures. The Split Sample Method (SSM) corrects for regression effects by using two performance measures, one as a measure of skill and the second one to calculate overestimation (Klayman, Soll, González-Vallejo, & Barlas, 1999).[10] If the two performance measures are parallel measures and the error components of these tests are uncorrelated, the SSM does correct for regression effects. Intuitively, this means that if students who have bad luck on the test which is used as the measure for skill are equally likely to have bad luck on the test which is used to determine their overestimation, there is no artificially negative relationship between performance and overestimation. However, there is still attenuation bias if the second performance indicator measures skill with random error.

Consider that there are two parallel performance measures (e.g. two test scores), p and p_0. Both performance measures are determined by skill and random errors with zero mean ε_p and ε_{p0}, respectively:

$$p = s + \varepsilon_p$$

$$p_0 = s + \varepsilon_{p0}$$

The SSM relies on the existence of two parallel measures. This means that the true skill enters equally into both performance measures that therefore have the same error variance, i.e. $Var(\varepsilon_p) = Var(\varepsilon_{p0})$ (Bollen, 1989). The crucial assumption for the SSM to correct for regression effects is that the errors ε_p and ε_{p0} are uncorrelated. This assumption can be problematic if there are common factors other than skill which influence both performance measures. Consider, for example, that p and p_0 are two parts of the same test. There might be many factors during the test (e.g. sickness, distracting seat neighbors) which will influence the performance on

10. This method was first used in an experiment where participants had to make a number of binary guesses and state their confidence in each guess in the range from 50 – 100 percent. In this context overestimation is the average confidence minus the average proportion correct. This method is called Split Sample Method, because in this context all items are randomly split in half and confidence and proportion correct are calculated for each half separately. Thus, proportion correct of the one subsample is used as p while the other subsample is used as p_0.

both test parts. In general, this assumption is therefore more plausible when the time between the two performance tests is considerably long and the measurement situation is different (Bollen 1989).

To show how the SSM corrects for regression effects but not for attenuation bias we can rewrite $\hat{\beta}_{SSM}$ as follows:

$$\hat{\beta}_{SSM} = \frac{Cov(p_0, oe)}{Var(p_0)} = \frac{Cov(s + \varepsilon_{p0}, oc - \varepsilon_p)}{Var(s) + Var(\varepsilon_{p0})}$$

$$= \frac{Cov\,(s, oc)}{Var(s) + Var(\varepsilon_{p0})}$$

Note that the covariance of p_0 and oe is equal to the covariance of s and oc if ε_{p0} and ε_p are uncorrelated and therefore the second term of Equation (6) disappears.[11] However, $Var(\varepsilon_{p0})$ still attenuates $\hat{\beta}_{SSM}$. This bias of $\hat{\beta}_{SSM}$ is proportional to β_s. The bias is positive if β_s<0 as suggested by the DK effect; the estimated $\hat{\beta}_{SSM}$ is thus closer to zero than the true β_s.

2.3.2 Reliability Adjustment

Attenuation bias drives the coefficient towards zero because $Var(\varepsilon_{p0})$ enters the denominator and reduces the absolute size of the estimated coefficient (see Equation (9)). One way to correct for this bias is reliability adjustment (Bollen, 1989). Test reliability (*r*) is defined as variance of skill divided by variance of performance:

$$r = \frac{Var(s)}{Var(p)} = \frac{Var(s)}{Var(s) + Var(\varepsilon_p)}$$

The reliability adjustment corrects for attenuation bias by dividing the coefficient by an estimate of test reliability. Intuitively, this procedure magnifies the coefficient by the factor for which it has been attenuated. Ehrlinger et al. (2008) estimated r by using the test-retest correlation, i.e. the correlation of the two test scores p and p_0. However, the test-retest correlation is only a correct measure of the test-reliability if both p and p_0 are parallel measures. This means that under the assumptions laid out for the SSM, the test-retest correlation leads to an unbiased estimate of the test reliability (Bollen, 1989).

11. This is because Cov(ε_{p0}, ε_p) = 0 if ε_{p0} and ε_p are uncorrelated.

One can see that reliability adjustment does correct for attenuation bias but increases the bias due to regression effects when we rewrite the reliability adjusted OLS estimate, $\hat{\beta}_{RA}$ as follows:

$$\hat{\beta}_{RA} = \frac{\hat{\beta}_{OLS}}{\hat{r}} = \left(\frac{Cov(s, oc)}{Var(s) + Var(\varepsilon_p)} - \frac{Var(\varepsilon_p)}{Var(s) + Var(\varepsilon_p)}\right) * \frac{Var(s) + Var(\varepsilon_p)}{Var(s)}$$

$$\hat{\beta}_{RA} = \frac{Cov(s, oc)}{Var(s)} - \frac{Var(\varepsilon_p)}{Var(s)}$$

When multiplying the coefficient by $\frac{Var(s)+Var(\varepsilon_p)}{Var(s)}$, which is equivalent to dividing it by the estimate of the test reliability, $Var(s) + Var(\varepsilon_p)$ cancel out in the numerator and denominator of both terms and we are left with $Var(s)$ in the denominator. Note that the first term is equal to an unbiased estimator of β_s while the second term is larger than the second term in Equation (6), which means that the overall bias of $\hat{\beta}_{SSM}$ is negative.

2.4 The Instrumental Variable Method

2.4.1 How the Instrumental Variable Method Works

The Instrumental Variable (IV) method is a popular bias correction method in economics (Angrist & Krueger, 2001) and epidemiology (Greenland, 2000).[12] It is a method to estimate causal effects of a variable which is endogenous, i.e. it is correlated with the error term of the equation, either due to reverse causality, omitted variable bias, or measurement error. The IV method uses additional variables, called instruments, to correct for the bias of the endogenous variable. These instrumental variables have to meet two requirements. First, they need to be correlated with the endogenous variable (relevance assumption). Second, they need to be uncorrelated with the error term of the main equation (exogeneity assumption).

For testing the DK effect a potential instrument is a second performance measure p_0. For a consistent estimate of β_s, this instrument needs to be correlated with the

12. See Angrist and Krueger (2001) for an intuitive and Wooldridge (2002) for a more technical introduction into the IV method.

original performance measure. This is the case when the SSM assumptions hold because both performance measures are partly determined by skill (see Equation (7)). In practice this condition is not critical and can easily be tested.[13] Moreover, it needs to be uncorrelated with the error term of the main equation, i.e. it should be uncorrelated with both ε_p and u_{oc}.[14] This assumption is also necessary for the SSM (see: Section 2.3.1). This second assumption is in practice more critical because it cannot be tested since ε_p and u_{oc} are unobserved. For this reason, it is important to plausibly argue that the instrument is not correlated with the error term of the main equation. This is more realistic when p and p_0 have been measured at different times and in different situations.

When the assumptions hold we can get a consistent estimate of β_s using a two stage procedure: In the first stage, the endogenous variable p is regressed on the instrumental variable p_0 :

$$p = \tau + \beta_{FS} p_0 + \omega$$

The estimated coefficient of the first stage, $\hat{\beta}_{FS}$, is equal to the correlation between the endogenous variable and the instrument. In the second stage, overestimation is then regressed on the predicted values of performance ($\hat{p}$) from the first stage instead of p:

$$oe = \alpha + \beta_{IV} \hat{p} + \epsilon$$

The IV estimator $\hat{\beta}_{IV}$ is equivalent to the estimator of the instrument in the main regression, which is called the reduced form estimate $\hat{\beta}_{RF}$, divided by the coefficient of the first stage $\hat{\beta}_{FS}$ (Wooldridge, 2002). This way of writing the IV estimator gives an insight into the way in which the IV method works: Because the instrument is uncorrelated with the error term of the main equation (see exogeneity assumption) $\hat{\beta}_{RF}$ is an unbiased estimate of the effect of the instrument on the dependent variable. By dividing $\hat{\beta}_{RF}$ by $\hat{\beta}_{FS}$ - the correlation between the endogenous variable and the instrument (see relevance assumption) - the size of the reduced form estimate is

13. In general, higher correlation between the instrument and the endogenous variable is preferred. See Murray (2006) for the problem of weak instruments.
14. Recall that the error term of the main equation is equal to $u_{oc} + \varepsilon_p - \beta_s \varepsilon_p$ (see Footnote 9). Therefore the instrument is not correlated with the error term of the main equation if it is uncorrelated with ε_p and u_{oc}.

adjusted to be equivalent to the effect of the endogenous variable (Angrist and Krueger 2001). To see how the IV method relates to the SSM and reliability adjustment, note that the reduced form estimate is the same as the SSM estimate $\hat{\beta}_{SSM}$ and the coefficient of the first stage is the same as the reliability estimate $\hat{r}$ (i.e. the correlation between p_0 and p). Hence, the IV method combines the two bias correction methods:

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{RF}}{\hat{\beta}_{FS}} = \frac{\hat{\beta}_{SSM}}{\hat{r}}$$

$$\hat{\beta}_{IV} = \frac{Cov\,(s, oc)}{Var(s) + Var(\varepsilon_{p0})} * \frac{Var(p)}{Var(s)}$$

$$\hat{\beta}_{IV} = \frac{Cov\,(s, oc)}{Var(s)}$$

When we divide the SSM estimator by test reliability (i.e. multiply by the inverse), $Var(p)$ and $Var(s) + Var(\varepsilon_{p0})$ cancel out, so we are left with $\frac{Cov\,(s,oc)}{Var(s)}$ which equals the unbiased estimator of β_s. This shows that, when the SSM assumptions hold, the IV method corrects for regression effects as well as attenuation bias.[15]

But what about $\hat{\alpha}$? In Section 2.2.1, we assumed that skill is strictly exogenous in Equation (1) and a hypothetical regression of overconfidence on skill would therefore give unbiased estimates of α and β_s. Thus, the estimate of the constant in this regression is equal to:[16]

$$\hat{\alpha} = \overline{oc} - \hat{\beta}_s \bar{s}$$

The estimate of the constant in the second stage of the IV regression (see Equation (13)) will therefore be equal to:

$$\hat{\alpha}_{IV} = \overline{oe} - \hat{\beta}_{IV} \bar{\hat{p}}$$

15. Yet another way to show that the IV estimator is consistent is by rewriting the estimator as $\frac{Cov(oe,p)}{Cov(p,p_0)}$ see: Angrist and Krueger (2001). It is then straightforward to show that this term is equal to $\frac{Cov(s,oc)}{Var(s)}$ if $Cov(\varepsilon_p, \varepsilon_{p0}) = 0$.
16. Recall that the constant of a binary regression is equal to $\bar{y} - \hat{\beta}\bar{x}$.

Note that when using OLS, the average of the predicted values is always equal to the average of the actual values and therefore $\bar{\hat{p}} = \bar{p}$. From Equations (14) and (15), it is straightforward to see that, because $\hat{\beta}_{IV}$ is equivalent to $\hat{\beta}_S$, $\hat{\alpha}_{IV}$ will be a consistent estimate of α if mean overconfidence is equal to mean overestimation ($\overline{oe} = \overline{oc}$) and mean skill is equal to mean predicted performance ($\bar{\hat{p}} = \bar{s}$). Both of which hold when ε_p has zero mean (see Section 2.2.2). The constant in the IV regression $\hat{\alpha}_{IV}$ is therefore a consistent estimate of α.

2.4.2 Potential Biases of the IV Method

So far we have assumed that expected performance is equal to skill plus overconfidence. When individuals are asked about their expected performance *after* the actual performance, individuals might have some idea on their luck ε_p on the test and adjust their forecast by a some adjustment factor ε_{exp}. In this case ε_{exp} would enter the overestimation measure and therefore become part of the error term of the main equation. However, in order for the exogeneity assumption to hold, and the IV method to lead to a consistent estimate of β_s, the instrument needs to be uncorrelated with ε_p, u_{oc} and ε_{exp}. Whether this assumption holds has to be considered for each application separately.

To understand the direction of potential biases of $\hat{\alpha}$ consider a situation in which ε_p has mean c, where $c \neq 0$. Because ε_p is part of the overestimation and performance measure c will affect $\hat{\alpha}$ through $\overline{oe}$ and through $\bar{\hat{p}}$. The bias of the constant is equal to $-c - \hat{\beta}_{iv}c$. Intuitively, think of a situation in which the grader gave each test taker c additional points. These points will increase the average performance and decrease the average overestimation. The extra points in the overestimation measure decreases $\hat{\alpha}$ by c while the extra points in the performance measure affects $\hat{\alpha}$ by $-\hat{\beta}_{iv}$c. If, for example, $\hat{\beta}_{iv}$ is -0.5 and c = 2 the bias of the constant is equal to $-(2) - (-0.5) * (2)$ = -1.

The IV method does not lead to consistent estimates when we use relative performance measures. Recall that when skill is measured with random error the transformation of absolute into relative measures leads to a negative correlation between relative skill and the relative error component (see Section 2.2.4). When two absolute performance measures, both measuring skill with random error in absolute terms, are transformed into relative measures there will be an artificial positive correlation between the two relative error terms, i.e. $Cov(\varepsilon_{rp1}, \varepsilon_{rp0}) > 0$. Intuitively, think of the individuals at the top and bottom of the skill distribution.

The individual with the lowest percentile can only have positive errors on both relative performance measures, while the individual with the highest percentile can only have negative errors on both performance measures. Through this bounding of relative performance the errors are positively correlated even when the errors of the underlying absolute performance measures are uncorrelated. Because $Cov(\varepsilon_{rp1}, \varepsilon_{rp0}) > 0$, the exogeneity assumption is violated and IV does not give a consistent estimate of the DK effect.

2.4.3 Advantages of the IV Method

The main advantage of the IV method over first estimating $\hat{\beta}_{SSM}$ and then dividing it by estimates of the test reliability (Bollen, 1989) is that it relies on less restrictive assumptions. Recall that both correction methods, the SSM and reliability adjustment, rely on the existence of parallel tests. Tests are more likely to be parallel when they are of the same format. However, tests of the same format are also more likely to have correlated errors. The IV method just relies on the assumption that the performance which is used as instrument is correlated with the endogenous variable and uncorrelated with the error term of the main regression. This allows researchers to combine different kinds of performance measures such as, for example, the use of last year's math course grade as an instrument for math performance on a recent test. When performance measures are different, the performance error is also less likely to be correlated and hence the exogeneity assumption is more likely to be met. Moreover, the IV method allows for the use of multiple instruments. Using several performance measures as instruments will increase the precision of the estimates. Including multiple instruments also enables us to test the exogeneity assumption (see Murray (2006) for a discussion of these tests). Finally, a more practical advantage is that the IV method is implemented easily with all current statistical software which automatically adjusts the standard errors accordingly.

2.5 Conclusion

While many studies have provided evidence for a negative relationship between performance and overestimation this does not mean that there also is a negative relationship between skill and overconfidence. The observed pattern between performance and overestimation might be a statistical artifact caused by the fact that performance measures skill with some error. When this error is random, which is a standard assumption for absolute performance measures, all of the currently used estimation methods give biased estimates of the DK effect. Moreover, when using

relative measures like percentiles, as it is frequently done in this literature, the transformation of absolute measures with random errors into relative measures creates a negative correlation between skill and measurement error. This negative correlation further complicates the analysis.

This means that to this date there is no convincing empirical evidence about the existence of the DK effect. Many studies use relative measures and it is not straight forward is which directions these are biased (Burson et al., 2006; Ehrlinger et al., 2008; Kruger & Dunning, 1999; Kruger & Dunning, 2002; Ryvkin et al., 2012).

When using absolute measures the behavior of the overall bias becomes more tractable. The overall direction of the bias of studies which estimate the DK effect with OLS is not clear (Kruger & Dunning, 1999). Studies which use the reliability adjustment lead to an overestimation of the DK effect (Ehrlinger et al., 2008). Theoretically, estimating the DK effect with the SSM would give us a lower bound of the true effect size. Unfortunately, the only studies which used the SSM estimated the DK effect with relative measures and therefore the overall bias of these study is not clear (Krueger & Mueller, 2002; Kruger & Dunning, 2002).

In this chapter, we suggest an alternative estimation method: the use of a second performance measure as instrument for skill. We have shown that, when using absolute measures, the IV method corrects for the two potential biases. IV estimation is thus a useful tool for testing the DK effect with absolute measures, which allows to further investigate the relationship between skill and overconfidence.

Chapter 3
Skill and Overconfidence*

* This chapter is based on joined work with Jan Sauermann and Andries de Grip. We thank Christian Kerckhoffs and Alexander Vostroknutov for access to their courses. We further like to thank Thomas Dohmen, and participants at the Maastricht University DuHR -PhD seminar for helpful discussions and comments.

3.1 Introduction

Overconfidence has been used to explain, among other things, financial bubbles (Scheinkman & Xiong, 2003), excessive mergers and acquisitions of CEOs (Malmendier & Tate, 2005) and excess market entry of entrepreneurs (Camerer & Lovallo, 1999). While most economic studies have not specified the relationship between overconfidence and skill, Kruger and Dunning (1999) proposed that it is generally the low skilled that are most overconfident while the high skilled are, on average, more accurate. This pattern between skill and overconfidence is called the Dunning-Kruger effect (DK effect) (Dunning, 2011). Kruger and Dunning (1999) proposed a lack of metacognitive skills of the low skilled as explanation for the DK effect. The intuition of this explanation is that the skill that is necessary to perform well is often the same skill necessary to evaluate one's performance. Hence, the low skilled do not only lack the skill to perform well but also the skill to recognize their low performance and are therefore more overconfident. Independent of the reason for the DK effect, such a pattern between skill and overconfidence would also be of interest for many situations which economists study. In general, the notion that the least able also tend to be most overconfident suggests that especially those who might fail are prone to spend more efforts than they should do (Ferraro, 2010). The DK effect could therefore explain, among other things, high college dropout rates (Turner, 2004) and excess market entry of entrepreneurs (Camerer & Lovallo, 1999).

Our study builds on a number of studies which have shown that for many different tasks low performers usually vastly overestimate their performance while high performers are, on average, more accurate or even slightly underestimate their performance (Burson, Larrick, & Klayman, 2006; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999; Ryvkin, Krajč, & Ortmann, 2012). We have shown in Chapter 2 that the fact that performance measures skill with some error can lead to biases when estimating the DK effect.

In this chapter, we show that - in contrast to the currently used Split Sample methods and reliability adjustment – the DK effect can be estimated consistently with an instrumental variable (IV) approach. We illustrate this IV approach in the context of incentivized exam grade predictions using students' grade point average (GPA) as an instrument for their exam grade. As prospective overconfidence is more relevant for choices which economists study we look at predictions instead of postdictions of

performance.[17] Our empirical estimates support the DK effect. Low skilled students are, on average, overconfident while high skilled students are more accurate. These estimates are substantially lower than OLS estimates and the reliability adjustment but substantially higher than the Split Sample method suggest. The contribution of our paper is both empirical and methodological. To the best of our knowledge this is the first paper that consistently estimates the DK effect.

The remainder of the chapter is structured as follows: Section 3.2 discuss the model, key variables and potential challenges when estimating the DK effect. Section 3.3 describes the data. Section 3.4 shows the results and Section 3.5 concludes.

3.2 The Model

As in Chapter 2, we model overconfidence (oc) as a linear function of skill (s):

$$oc = \alpha + \beta_s s + u_{oc} \qquad (1)$$

Overconfidence is the sum of a constant term, α, and a variable component that depends on the individual's skill; u_{oc} is an idiosyncratic error term which captures individual differences in overconfidence that are unrelated to skill. The aim of this chapter is to estimate α and β_s. The DK effect suggests that skill is negatively related to overconfidence, i.e. that β_s is negative. Combining α and β_s shows estimates of the average overconfidence for different skill levels. The DK effect thus, predicts high positive level of overconfidence for the low skilled and low levels of overconfidence/underconfidence of the high skilled. This is tested against a null-hypothesis that there is no relationship between skill and overconfidence, i.e. that β_s is zero indicating that the high and low skilled are equally overconfident. If we could observe overconfidence and skill directly, and skill would be strictly exogenous in Equation (1), an OLS regression of overconfidence on skill would lead to unbiased estimates of α and β_s. Skill and overconfidence, however, are unobservable and researchers use performance on a test and overestimation of this performance as their respective measures.

We define all key variables, skill, performance, overconfidence and overestimation as we did in Chapter 2. We define skill straightforward as the ability in the relevant

17. In the psychological literature several studies on the DK effect use expectations of test results after the test. The statistical properties of the IV method, however, also apply for these postdictions of performance.

domain. The available measure of skill is performance (p) on a test which is the sum of skill and random measurement error (ε_p):

$$p = s + \varepsilon_p \tag{2}$$

We define overconfidence as the difference between the self-assessed skill level and the actual individual skill level and overestimation as the difference between expected and realized performance. Expected performance (p_{exp}) is equal to the sum of a person's actual skill and overconfidence:

$$p_{exp} = s + oc \tag{3}$$

When decomposing overestimation into its respective elements one can see that it is equal to the difference of oc and ε_p:

$$\begin{aligned} oe &= p_{exp} - p \\ oe &= (s + oc) - (s + \varepsilon_p) \\ oe &= oc - \varepsilon_p \end{aligned} \tag{4}$$

When we use overestimation as a measure of overconfidence, the exam error thus enters the error term of the equation:

$$oc = \alpha + \beta_s s + u_{oc} + \varepsilon_p \tag{5}$$

When we look at Equation (5) it becomes clear that using performance of the same test that is used to calculate overestimation as a measure for the individual skill level will cause regression effects because of ε_p. Performance is correlated with the error term because ε_p is included in the measured performance as well as overestimation. Further, if ε_p is random it will lead to attenuation bias (see Chapter 2).

This endogeneity problem can be solved by an instrumental variable (IV) approach. IV estimation can get a consistent estimate of β_s with the help of an additional variable which is correlated with the endogenous variable (relevance assumption) and uncorrelated with the error term of the main equation (exogeneity assumption). If these two assumptions are met, β_s can be estimated consistently with the IV approach. In this study, we use a second performance measure which is correlated with the first performance measure but arguably uncorrelated with u_{oc}, and ε_p.

When the error term has zero mean the IV method will also lead to a consistent estimate of α.

3.3 Data

Our sample consists of economics students of two second year bachelor courses, which were given in March and April 2013 at the School of Business and Economics (SBE) of Maastricht University in the Netherlands.[18] 91 percent of the students in our sample are in same economics Bachelor program and each course is a compulsory course of a different specialization of this program. The remaining 9 percent of students are from other Bachelor programs who took this course as an elective. No student took both courses, but 87 percent of the students in our estimation sample took the same eight compulsory courses in the first year of their study. Because Maastricht is close to the German border, the SBE has a large share of German students. In our estimation sample 50 percent of students are German and 30 percent are Dutch; 31 percent are female.

We elicited students' predictions of their exam grade with a questionnaire four weeks before the exam.[19] Grades are given on a scale from 0 (lowest) to 10 (highest) in course 1 and from 1 to 10 in course 2. For both courses, the minimal exam grade necessary to pass the course is 5.5. To ensure that students state their honest expectations we incentivized the exam grade predictions by holding a lottery draw where students could win in each course one of two gift vouchers worth €20 if their prediction was within a range of 0.25 points around their actual exam grade (see questionnaire in the appendix). Furthermore, students were assured that all information would be kept confidential. Information on actual grades was provided by the course coordinators, information on student characteristics and previous grades is taken from the administrative records.

In total 165 of 209 registered students filled out the questionnaire which means a response rate of 79 percent. The remaining 21 percent were not present during the distribution of the questionnaire in the class room; either because they missed the particular session or because they had already dropped out of the course. Table 3.2 shows the summary statistics for the estimation sample of students' predictions, actual grades, the resulting over- and underestimation as well as students' GPA at

18. See Chapter 4 for more information on the institutional background.
19. We also elicited students' expectation about the percentile of their exam grade and their participation grade. We do use these predictions to test the DK effect because we do not have suitable instrument for participation grade problems of relative measures discussed in Chapter 2.

the end of the first year. On average students significantly overestimated their exam grade ($p = 0.004$).

Table 3.1 Predictions, Grades and Overestimation

	Obs.	Mean	S.D.	Min	0.25	0.50	0.75	Max
Predicted exam grade	153	7.22	0.85	4.50	6.50	7.00	8.00	9.25
Realized exam grade	153	6.85	1.93	0.00	5.75	7.00	8.00	10.00
Exam overestimation	153	0.37	1.70	-3.00	-0.75	0.25	1.50	6.20
GPA	153	7.17	1.15	4.34	6.37	7.17	8.08	9.38

Note: Based on estimation sample. Exam overestimation is equal to predicted minus realized exam grade.

In this study, we use the students' grade point average (GPA) as an instrument for their exam grade - the (endogenous) measure of their skill. The instrument needs to be correlated with the exam grade but uncorrelated with the error term of the main equation. We propose students' first year GPA, which is calculated as the weighted average of all grades at the end of the first year,[20] as an instrument for the following reasons. First, it is correlated with the exam grade because it is often similar skills that determine grades in different courses. Second, because the last grade of the first year GPA was graded eight months before the exam students were asked to predict, the GPA is arguably uncorrelated with the exam error ε_p. However, the GPA might be correlated with u_{oc} if there are other factors which are correlated with skill as well as overconfidence. Student gender or nationality, for example, might be correlated with skill and overconfidence. To account for this possibility we estimate Equation (5) including dummies for gender, student nationality, course dummies and a dummy indicating whether the exam was taken in the second attempt.

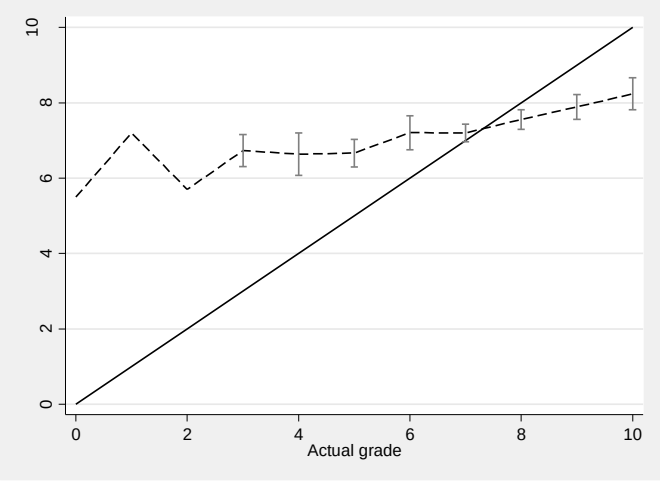
20. The GPA is a weighted average (by ECTS points) of all graded components available at the end of the academic year 2011/2012. The same data is used in Chapter 5. For most of the students the GPA measure consist of eight regular courses (6.5 ECTS) and two skills courses (3 ECTS) which are compulsory in the first year of the economics Bachelor.

3.4 Results

3.4.1 Main Results

Figure 3.1 plots the average exam predictions against the actual exam grades. If all individuals would have perfect foresight about their exam grades, the relation between predicted and actual grades would be shown by 45-degree line (solid line). The figure shows the typical pattern of many DK effect studies: those with lower grades vastly overestimate their exam grade while those with higher grades slightly underestimate their exam grades. However, as discussed in Section 3.2, the relationship between performance and overestimation shown in Figure 3.1 is biased due to regression effects and attenuation bias.

Figure 3.1 Actual versus Predicted Exam Grades



Note: The figure shows predicted exam grades against actual exam grades. Caps show the 95% confidence interval of the predicted exam grades.

Table 3.2 shows the first and second stages of an IV regression with students' overestimation as dependent variable, students' exam grade as endogenous variable and students' GPA as instrument. Column (1) and (3) show the first stages without and with additional control variables respectively. Unsurprisingly, past GPA is highly correlated with a student's current exam grade. Students who performed well in previous courses are very likely to perform well on the current exam as well. If we control for background information of the student, the estimated coefficient is slightly smaller, yet not statistically different. The F-tests of the excluded instrument

are large. Columns (2) and (4) of Table 3.2 shows the estimated coefficients of the second stage without and with additional controls respectively. Both estimates are negative and highly significant. The coefficient of our preferred specification in Column (4) suggests that a one grade point increase in skill reduces overconfidence by 0.6 grade points.

These results provide evidence for the DK effect: the negative coefficient of the (predicted) exam grade shows that students with higher skills are, on average, less overconfident while students with lower skills are more overconfident. We can further use the predicted exam grades, our unbiased measure of skill, and the respective estimates of α and β_s from Column (4) to demonstrate that the DK effect holds in our sample: The predicted overconfidence of the student of the 10th percentile of the skill distribution (predicted exam grade = 5.00) is equal to 2.18 (5.18 -0.60*5.00) while the predicted overconfidence of a student of the 90th percentile of the skill distribution (predicted exam grade = 8.42) is equal to 0.12 (5.18 -0.60*8.42). In line with the DK effect, the low skilled are very overconfident while the high skilled are more accurate.

Table 3.2 IV Estimates of the DK Effect

	(1)	(2)	(3)	(4)
	First Stage	Second Stage	First Stage	Second Stage
	Exam grade	Overestimation	Exam grade	Overestimation
Exam grade		-0.5539***		-0.5964***
		(0.080)		(0.067)
GPA	0.9654***		0.9710***	
	(0.089)		(0.109)	
Constant	-0.0663	4.1632***	-0.5737	5.1781***
	(0.696)	(0.601)	(0.823)	(0.520)
F-test excluded instrument		118.9		79.1
Observations	153	153	153	153
R-squared	0.329	0.733	0.408	0.798

Note: Standard errors clustered at section level in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Additional controls include dummy variables for female, German, Dutch, field of study economics, course 2, resit exam.

3.4.2 Robustness Checks

The results from Table 3.2 are remarkably robust. Estimating the DK effect separately for course 1 and 2, including the grades of all eight compulsory courses separately or jointly leads to qualitatively similar results (not shown). There is one

potential concerns about the validity of the instrument we used. It might be that overconfidence directly influences a student's grade. If students who are overconfident behave differently in the classroom the teacher and grader might confuse this confidence for skill and therefore (subconsciously) give higher grades to more overconfident students. This is possible if the grader, who is in many cases the same person as the teacher, knows the students because at the SBE students are expected to write their names on their exams.[21] If this is the case, overconfident students would appear to be more skilled and less overconfident and our estimates would be biased. To test whether our estimates are driven by the effect of overconfidence on grades we redid the analysis using a performance measure which was graded objectively as instrument: The grade of a first year course which only graded component is a machine graded multiple choice (MC) exam. Table 3.3 shows the analysis similar to columns (3) to (4) in Table 3.2 but with MC course grade instead of GPA as an instrument. Column (1) shows that MC course grade significantly predicts exam grade on the current course and the F-statistic is large. The point estimate reported in Colum (2) is remarkably close to the point estimate reported in Table 3.2. This result suggests that our results are not driven by the effect of overconfidence on grades.

Table 3.3 Robustness, Multiple Choice Course Grade as Instrument

	(1) First Stage Exam grade	(2) Second Stage Overestimation
Exam grade		-0.6198***
		(0.076)
Multiple choice course grade	0.8089***	
	(0.093)	
Constant	1.4953*	4.6692***
	(0.716)	(0.529)
F-test excluded instrument		76.3
Observations	135	135
R-squared	0.346	0.811

Note: Standard errors clustered at section level in parentheses; *** p<0.01, ** p<0.05, * p<0.1. Additional controls are as in Columns (3) and (4) of Table 3.2.

21. See Chapter 4 for more details about the examination and grading procedure at the SBE.

3.4.3 Contrast Findings to Current Methods

How do the IV estimates presented in Section 3.4.1 compare to the estimates we would have got with the estimation methods currently used in the literature? Column (1) of Table 3.4 shows the IV estimates from our preferred specification of Column (4) Table 3.2 which we will use as baseline. Column (2) shows the results of an OLS estimation of overestimation on performance and additional controls. The OLS coefficient is about 35 percent larger than the IV coefficient which shows that in our sample OLS substantially overestimates the magnitude of the relationship between skill and overconfidence. This shows that in this setting the regression effects bias is larger than the attenuation bias.

Unfortunately we do not have data of the students' performance on a second comparable exam. With data on such an exam we could have calculated the test-retest correlation and then apply the reliability adjustment method (Ehrlinger, et al., 2008). Further, we could have applied the Split Sample method by using the comparable second exam grade as our measure of skill (Klayman, et al., 1999). However, we can simulate the magnitude of the reliability adjusted OLS estimator and the Split Sample estimator if we assume some plausible value for a test-retest correlation between the two exam grades. Values for test-retest correlation in DK effect studies typically range from 0.50 to 0.80. (Burson, et al., 2006; Ehrlinger, et al., 2008; Klayman, et al., 1999; Krueger & Mueller, 2002). We will use 0.80, an estimate most favorable to other methods used in the literature, as plausible value for a test-retest correlation. Column (3) of Table 3.4 shows the simulated reliability adjusted OLS estimator, which is equal to the OLS estimator divided by the hypothetical test-retest reliability 0.80, and the accordingly adjusted estimate the constant (see: Ehrlinger, et al., 2008). Note that, if measurement error is random, the Split Sample estimator only suffers attenuation bias and test-retest reliability is an indicator of the degree of this attenuation (see: Chapter 2). Hence, we can simulate the size of the Split Sample estimator by multiplying the IV estimator by the hypothetical test-retest reliability. Column (4) of Table 3.4 shows simulated coefficient and the accordingly adjusted constant.[22] The OLS estimates and the simulations illustrate that the different estimation techniques can lead to substantially different estimates. While in our sample all methods would have supported the existence of the DK effect they differ vastly in magnitude. Compared to the consistent IV estimator, the OLS estimator substantially overestimates the DK

22. The formula to calculate the new constants for the reliability adjusted OLS and the Split Sample method is: $\overline{oe} - \hat{\beta}\overline{p}$, where $\hat{\beta}$ is the respectively adjusted coefficient. The average overestimation of the estimation sample is 0.368 and the average performance of the estimation sample is 6.852.

effect. As we would expect, the reliability adjusted OLS leads to an even larger overestimation of the DK effect because it only corrects for attenuation bias and even aggravates regression effects. Further, the Split Sample method, which only corrects for regression effects, underestimates the DK effect. It is straightforward to see that for more unreliable performance measures these differences would be even larger.

Table 3.4 Comparison of Different Statistical Methods to Estimate the DK Effect

	(1) Estimated IV Overestimation	(2) Estimated OLS Overestimation	(3) Simulated Ra. OLS Overestimation	(4) Simulated SSM Overestimation
Exam grade	-0.5964***	-0.8077***	-1.0096	-0.4771
	(0.067)	(0.032)		
Constant	5.1781***	6.5590***	7.2858	3.6371
	(0.520)	(0.311)		
Observations	153	153	153	153
R-squared	0.798	0.849		

Note: Standard errors clustered at section level in parentheses; *** p<0.01, ** p<0.05, * p<0.1. Column (1) is equal to column (4) in Table 3.3 and shows estimates of IV regression where first year GPA is used as an instrument for current exam grade. Column (2) shows estimates of an OLS regression of overestimation on exam grade. Additional controls for Column (1) and (2) are the same as in Table 3.2 column (4). Columns (3) and (4) are simulated: The coefficient of in column (3) is equal to the coefficient in column (2) divided by 0.80. The coefficient in column (4) is equal to the coefficient in column (1) multiplied by 0.80. Constants in Columns (3) and (4) are calculated as described in Footnote 22.

3.5 Conclusion

It is difficult to test for the existence of the DK effect because measurement error causes regression effects and attenuation bias. We are the first to consider both of these biases simultaneously by estimating the DK effect with an IV approach. In the context of students' exam grade predictions we find evidence for the DK effect: Low skilled students are overconfident while high skilled students are more accurate. This finding is robust to a number of potential concerns. We further show that estimating the DK effect with the methods currently used in the literature leads to substantially different estimates.

Future research will have to show whether the DK effect also holds for different tasks and in different population. If the DK effect, however, is a robust

psychological regularity this can inform the study of a number of situations in which accurate human judgment and decisions are crucial. The existence of the DK effect would suggest that even low levels of overconfidence on average in a given population hides substantial heterogeneity with the low skilled, those who often most benefit from accurate self-assessment, being most overconfident.

3.6 Appendix

Questionnaire

The only difference between the questionnaires for course 1 and 2 is that the prediction for the participation grade was only incentivized for course 2. Therefore students could win up to two vouchers in course 1 worth €40 and up to three vouchers in course 2 worth €60. Differences between the questionnaires are indicated with "→*only for course 2*".

Dear student,

I am Jan Feld, PhD student in Economics at the School of Business and Economics. My research concerns the relation between grade expectations and realised grades.

I would like to ask you for your expectations of your grade in the [course name] exam and your participation grade. Please give your best estimates. You can enter three lotteries if your estimates are close to your actual results. In each lottery you can win one of [two/three] VVV vouchers worth €20. In total, you can win VVV vouchers of [€40/€60].

At the end of the survey, you will be asked to enter your student ID. The ID is required to compare your estimates with your actual results. If you win one of the lotteries, the ID will be used to look up your email so that I can inform you about your win.

I will treat this information confidentially and ensure your anonymity. No individual information will be passed on to anybody (not even your tutor or course coordinator). I will also not report any information which can be used to identify you.

If you have any questions, please feel free to contact me via: j.feld@maastrichtuniversity.nl

Thank you for your cooperation!

Jan Feld

This is how the lotteries are going to work:

Lottery 1: If your exam grade (in your first attempt) is within 0.25 points of your expected grade, you enter a lottery in which two winners are randomly drawn. If you do not attend the first sit, your second sit grade is considered for the lottery. Each winner will receive a VVV voucher worth €20.

Lottery 2: I calculate the actual percentile of your exam grade compared to the exam grades of the first attempts of all students in this course. If your final exam grade is in your expected percentile range, you enter a lottery in which two winners are randomly drawn. Each winner will receive a VVV voucher worth €20.

[Lottery 3: If your actual participation grade is within 0.25 points of your expected participation grade, you enter a lottery in which we randomly draw two winners, who will receive a VVV voucher worth €20.] →*only for course 2*

Questionnaire Grade Expectations - Course [course name]

****

Which grade do you expect to get in the exam of the course [course name]?

If you do NOT intend to attend the first sit, please state your expectations for the second sit (resit).

I expect to get a __ __.__ __ in the exam. [0.00-10.00]

Please indicate in which percentile range you expect your exam grade to be in?

The percentile shows the percentage of students in this course which have a lower exam grade (in their first attempt) than you. High values mean high exam grades compared to the exam grades of the other students in this course.

Please mark your expected percentile range with an X.

	1%-10%	11%-20%	21%-30%	31%-40%	41%-50%	51%-60%	61%-70%	71%-80%	81%-90%	91%-100%
Your percentile:										
	Worst 10%									*Best 10%*

Which participation grade do you expect to get in this course?

[Please state your guess rounded to the next quarter point so that it ends with .00, .25, .50 or .75.] → only for course 2

I expect to get a__ __ .__ __as participation grade. [0.00-10.00]

Do you consider failing on purpose in the first sit of the exam in this course – either by not attending or by handing in an incomplete exam – in order to get a higher grade in the second sit?

Yes ☐ No ☐

What is your gender?

Male ☐ Female ☐

What is your student ID?

ID_____

Please fold this page in half after filling it out.

Chapter 4
Endophilia or Exophobia: Beyond Discrimination*

* This chapter is based on joined work with Nicolás Salamanca and Daniel Hamermesh. We thank Jeannette Hommes, Ad van Iterson and Caroline Kortbeek for their assistance in making this experiment possible. Eric Bonsang, George Borjas, Deborah Cobb-Clark, Andries de Grip, Thomas Dohmen, Hannah Ebin, Matthew Embrey, Ilyana Kuziemko, Corinne Low, Arjan Non, Christopher Parsons, Joseph Price, Stephen Trejo, participants in seminars at a number of universities and institutes, and especially Leigh Linden provided very helpful comments. The Board of Examiners of the School of Business and Economics at Maastricht University formally approved this project.

Although we could not perceive our own in-groups excepting as they contrast to out-groups, still the in-groups are psychologically primary. Hostility toward out-groups helps strengthen our sense of belonging, but it is not required. [Allport, 1954]

4.1 Introduction

Economists have studied labor-market discrimination at least since Becker (1957). Differences in labor-market and other outcomes by race, gender, ethnicity, religion, weight, height, appearance and other characteristics have been examined in immense detail, over time and in many economies. The focus has, however, been nearly exclusively on measuring differences in outcomes between groups, under the assumption that the "majority" group's outcome is the norm while the "minority" group is discriminated against. But since the only concept that is measured is a difference, it could just as easily be that the majority group is favored while the minority group's outcome is the norm.

The possibility that we are measuring the extent of favoritism rather than discrimination has been pointed out by Goldberg (1982) and by Cain (1986) in his survey; but beyond that the issue appears to have been completely neglected, including by the more recent *Handbook* surveys of the literature on discrimination (Altonji and Blank, 1999; Fryer, 2011). Once we recognize that favoritism need not be the obverse of discrimination, the importance of studying preferences for favoritism/discrimination increases. Although the distribution of discriminating agents' tastes underlay Becker's theory, in most empirical research the demand side—the behavior of discriminatory agents—has not been studied explicitly. Only recently has there been even a small upwelling of interest in examining their behavior and its impacts on outcomes.[23] These studies typically consider how agents' behavior toward those who match them along some dimension differs from their behavior toward those who do not match them, again only estimating relative differences. Even then, most of these studies have looked only at averages, and none has combined this with the analysis of the distributions of preferences.

Here we discuss the results of a field experiment that allows us to identify separately the means of favoritism and discrimination, as well as their distributions. The key to

23. See Price and Wolfers (2010) and Parsons *et al* (2011) for evidence from professional sports; Fong and Luttmer (2009) on charitable giving; Dee (2005), Lavy (2008), Hinnerich *et al* (2011), and Hanna and Linden (2012) for examinations of education; Cardoso and Winter-Ebmer (2010) and Giuliano *et al* (2011) on wages and hiring; Baguës and Esteve-Volart (2010) on parliamentary elections; and Dillingham *et al* (1994), Donald and Hamermesh (2006) and Abrevaya and Hamermesh (2012) for studies of economists' behavior.

doing this that, instead of measuring differences in outcomes *between* groups, we compare outcomes of members of the *same* group with and without visible characteristics that reveal to which group they belong.[24] In the context of our experiment, we do this by randomly revealing or concealing names on students' final exams, and thus randomly allowing or not allowing graders to infer the gender and nationality of the students. Because of the random assignment, students without visible names on their exams have on average the same observable and unobservable characteristics as students with visible names on their exams. Students without visible names thus serve as a neutral baseline to identify discriminatory preferences. Differences from this baseline can be entirely attributed to the presence of the name—and by inference to favoritism/discrimination.[25] Hence, we have evidence for favoritism if members of a group are treated better when their names are visible. Conversely, we can infer the presence of discrimination if members of a group are treated worse when their names are visible. We focus specifically on favoritism/discrimination by gender and nationality, but this method could be applied to any of the groups that have been studied in this immense literature.

To distinguish clearly the *who* and the *how* in discrimination, we introduce four terms: Endophilia, endophobia, exophilia, and exophobia. The prefix *endo* refers to preferences towards people like oneself, the prefix *exo* to people unlike oneself. The suffixes *philia* and *phobia* refer to favoritism to discrimination. Hence, *endophilia* denotes preferences for member of one's own group, while *exophobia* denotes preferences against members of other groups. One can also imagine, however, that some agents prefer members of other groups—are *exophilic*, while other agents are *endophobic*—discriminate against people like themselves.

24. A number of studies (e.g., Goldin and Rouse, 2000, Burgess and Greaves, 2013) have focused on "blindness" in quasi-experimental situations to infer the extent of discrimination (or favoritism, since neither study could distinguish between these).
25. The only experiments like ours were conducted in laboratories (Fershtman *et al*, 2005; Ahmed, 2007). The latter had artificially-designated in- and out-groups; the former dealt with nationalities but was based on statements by students on how they would behave in a trust game. While laboratory evidence is useful, as discussed by Levitt and List (2007) it suffers from a number of difficulties that can be addressed in field experiments.

4.2 Theoretical and Empirical Motivation

The importance of the distinction between favoritism and discrimination can be seen both theoretically and empirically. Our theoretical work is a generalization of Becker's (1957) theory of discrimination and Goldberg's (1982) alteration of it. Goldberg adapted Becker's model to show that if favoritism toward one's own group drives observed, apparently discriminatory wage differentials, these differentials can persist in a competitive market. He reached this conclusion by assuming that employers have favoring *instead of* the discriminatory preferences as in Becker (1957). Employers can, however, have both discriminatory *and* favoring preferences. We extend Becker's model to show that if both preferences are present, the intergroup wage differential will not only depend on the distributions of favoritism and discrimination but also on the relationship between their distributions.

Assume, as Becker does, that all employers are White and that there is a fixed labor force, some fractions of which are White and Black. Let employers have endophilic and exophobic preferences simultaneously, so that we can characterize a typical employer's utility as:

$$U = Q(L_w + L_b) - W_w L_w - W_b L_b + eL_w - xL_b \ ,$$

where we assume as usual that White and Black workers are perfect substitutes in the production function Q for the good sold at a constant price of unity. This utility function implies that employers obtain or forgo a fixed amount of utility when they hire Whites or Blacks according to their preferences.[26] Analogous to both Becker's and Goldberg's models, employers here will hire Whites or Blacks depending on whether $W_w - W_b$ is smaller or larger than $\delta = e + x$, the relative preference for Whites over Blacks.

In equilibrium, Blacks will be employed by employers with lower values of δ and Whites will be employed by the remaining employers. Knowing this, we can find a simple expression that implicitly identifies the preferences of the marginal employer. Assume that the distribution of endophilia is $f(e)$ and of exophobia is $h(x)$ across the population of jobs on offer, and denote the relative preference of the marginal

26. Goldberg and Becker model discrimination and favoritism as wage premia and discounts rather than as a fixed utility gain or loss. Although more complex, the model here and the general intuition do not change if we choose to model discrimination and favoritism as they do.

employer, for whom $W_w - W_b = e + x$, as δ^*. Then in long-run equilibrium the share of Blacks in the economy determines δ^*:

$$\frac{L_b}{L_b + L_w} = \int_{-\infty}^{\delta^*} g(\delta) d\delta,$$

where $g(\delta)$ is the density function of δ which, in general, will depend on the densities f and h and their relation. If $e \sim N(\mu_e, \sigma_e^2)$ and $x \sim N(\mu_x, \sigma_x^2)$:

$$\frac{L_b}{L_b + L_w} = \Phi\left(\frac{\delta^* - \mu_e - \mu_x}{\sqrt{\sigma_e^2 + \sigma_x^2 + 2\sigma_e\sigma_x\rho}}\right),$$

where Φ is the cumulative standard normal distribution and ρ is the correlation between endophilic and exophobic preferences in the population of jobs being offered. It is easy to see that, keeping the shares of Blacks and Whites and the means and variances of the densities f and h constant, an increase in ρ will result in a marginal discriminator with weaker relative preferences for Whites over Blacks. The increase in ρ increases the variance of δ, the sum of preferences, while keeping its mean constant. This leads to a distribution of relative preferences for Whites with more extreme values; more employers now have an extremely strong and extremely weak relative preference for Whites. Because Blacks are hired by employers who are most favorable to them, the marginal discriminator now has a lower relative preference for Whites. A more positive correlation of endophilia and exophobia thus leads to a decrease in the absolute value of the equilibrium Black-White wage gap.[27] In other words, for the same means and variances of endophilic and exophobic preferences, and holding constant the share of Black workers, the wage gap is smaller if the most bigoted (against Blacks) employers are also those who favor Whites most.

By allowing agents to have endophilic and exophobic preferences at the same time, our model becomes a more general version of both Becker's and Goldberg's, effectively nesting both cases.[28] The model predicts that the wage gap increases if employers become more exophobic (as in Becker) and as they become more endophilic (as in Goldberg). By allowing endophilia and exophobia to co-exist,

27. See Charles and Guryan (2008) for a discussion of the empirical importance of the marginal discriminator.
28. If we assume that endophilia is non-existent, the model reverts to Becker's; if we assume that exophobia is non-existent, it reverts to Goldberg's.

however, the model introduces an additional force that can shape market outcomes, the correlation of the two types of preferences.

The concepts of endophilia, exophobia, and their correlation can be measured, albeit imperfectly, in the real world. Beginning in 1996, and biennially except in 2002, the U.S. General Social Survey has asked questions, "In general, how close do you feel to Whites [Blacks]?" with answers on a nine-point scale ranging from 9 = very close to 1 = not close at all. Table 4.1 describes these data, separating answers by Whites and Blacks, and pooling 1996-2000 as an early period, 2004-2006 as a later period. (We exclude the 2008 and 2010 data because the campaign and election of President Obama may have altered expressed preferences.) Several things stand out: 1) Unsurprisingly, expressed closeness to one's own group exceeds that to the other group; 2) While Whites' closeness to other Whites changed little over this period, there was a very large increase in their expressed closeness to Blacks; 3) There are only small changes in Blacks' expressed closeness to either Whites or Blacks; and 4) The correlation between expressed closeness to one's own group and the other is positive and increased (significantly) between the two sub-periods. Implicitly, those who favor members of their own group more disfavor members of the other group less, or, in our terminology, there was an increasing negative correlation between endophilia and exophobia.

To illustrate how thinking about endophilia and exophobia jointly can add to our understanding of discriminatory outcomes, consider the implications of the GSS data for the evolution of the Black-White wage gap. Assume for simplicity that the share of Black workers remained constant and that all employers are White. In Table 4.1 we can see between 1996-2000 and 2004-2006 Whites' endophilia remained constant, Whites' exophobia (the negative of the measure in the Table 4.1) decreased, and the correlation between endophilia and exophobia decreased (became more negative).[29]

29. Their correlation decreases because Whites' closeness to Blacks, as reported in Table 4.1, is measuring exophilia, the opposite of exophobia.

Table 4.1 Endophilia and Exophobia in the U.S. General Social Survey, 1996-2006, 9-point scale

Time period:	1996-2000	2004-2006
WHITES		
Feel Close to Whites	7.060	6.966
	(0.031)	(0.038)
Feel Close to Blacks	5.121	5.494
	(0.032)	(0.039)
N	3,550	2,174
P	0.146	0.226
BLACKS		
Feel Close to Whites	5.799	5.945
	(0.084)	(0.106)
Feel Close to Blacks	7.547	7.685
	(0.079)	(0.093)
N	651	387
P	0.242	0.318

Note: In general, how close do you feel to …? not close at all = 1; very close = 9.

Becker's model (where only exophobia matters) predicts that the decline in exophobia shown in the table would decrease the wage gap. Goldberg's model (where only endophilia matters) predicts that the wage gap would remain constant. Our model captures one force that tended to decrease the gap—the reduction of exophobic preferences; and one that tended to increase it—the more negative correlation between endophilia and exophobia. Perhaps these opposite forces contributed to the constancy of the black-white earnings ratio over this period, although with so many other shocks over this short period attributing changes is difficult.

4.3 Constructing the Experiment

4.3.1 The Environment

To make the distinction between favoritism and discrimination empirically we set up a field experiment that we carried out during the final exam week in June 2012 at the School of Business and Economics (SBE) of Maastricht University in The Netherlands. The language of instruction throughout the SBE is English. This environment has a number of features that make it particularly appropriate for distinguishing between favoritism and discrimination. Partly because Maastricht is near the German border, the SBE has a large share of German students (51 percent) and academic staff (22 percent) mixed with Dutch and other nationalities. The student population is 36 percent female, and the academic staff is 28 percent female.[30] German students have a reputation for being more hard-working than Dutch and other students. These contrasts by nationality could potentially be the basis for discrimination/favoritism, although it is unclear *a priori* in which direction these will be.[31]

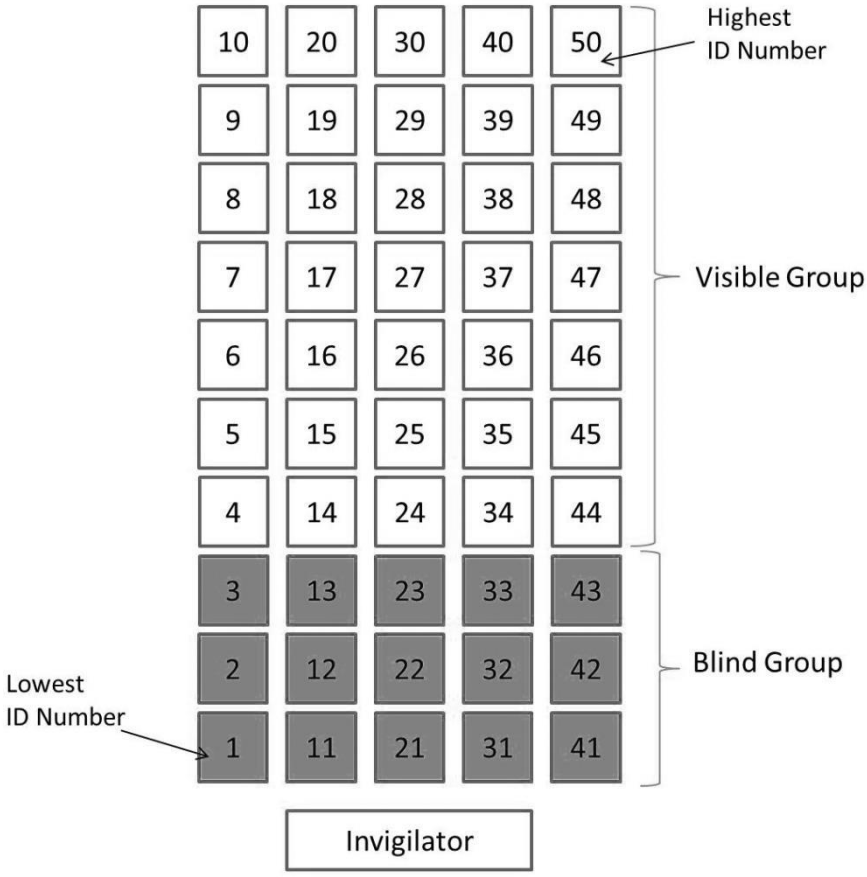
The grading of final exams, which we examine here, is a good setting for identifying discrimination/favoritism, because graders do not gain anything from favoring or disfavoring specific groups. Also, until the teaching period that we examine all students were required to write their names on their exams, enabling the graders to identify the students' gender and nationality.[32] Finally, and most important, this experiment has real-world consequences: The grades are important to students; also, much of the graders' jobs revolves around their role in scoring exams.

30. The SBE homepage (http://www.fdewb.unimaas.nl/miso/index.htm) provides these statistics for enrolled students in 2010 for nationality and 2012 for gender. Statistics about staff refer to full-time-equivalent academic staff in 2012 and are taken from the internal information system "Be Involved."
31. While it is often found that people favor (discriminate against) groups with same (different) characteristics, there are also situations in which the opposite is the case. One can, for example, think of many situations in which relative outcomes suggest that males are exophilic or endophobic (e.g., Donald and Hamermesh, 2006, although that study cannot distinguish between these two types of preferences).
32. The grader can infer the nationality and gender of the students when she sees the family name, even if she does not know the student, because Dutch and German names are quite distinct. To test this we asked 9 staff (5 German and 4 Dutch, of whom 5 were female) to guess the nationality and gender of 50 student names from our sample. We selected the student names block-randomly to reflect the nationality mix in our sample (19 German, 17 Dutch and 14 other nationalities, of whom 16 were female). The staff correctly identified the German names in 64 percent and Dutch names in 65 percent of cases, and they correctly guessed gender in 90 percent of the cases. On the other hand, graders may be more able to infer student gender than nationality from handwriting *per se*.

In the SBE written exams are administered in ten sessions spread over a week, with many courses giving their exams simultaneously. Students in all the courses assigned to each session take their exams together in a large conference hall filled with desks that are arranged in blocks of 5 columns and 10 rows.[33] To prevent cheating the location of each student's desk is predetermined by the Exams Office (the organization responsible for examination procedures). The desk assignment is based on student ID numbers, first by sorting them from lowest to highest within each block, and then filling in sequentially within each column from left to right.[34] Figure 4.1 illustrates the arrangement of desks in each block.

33. Exams in courses with more than 50 students are written in the same session in multiple blocks. Exams in courses with fewer than 50 students are either kept in one block or are combined with the exams in other courses. There are a few blocks that have as many as 12 rows.
34. Student IDs are assigned in ascending order based on the moment a prospective student contacts Studielink (the Dutch centralized system for university application; https://app.studielink.nl/front-office/). This means that earlier cohorts have lower-number IDs, and later cohorts and exchange students have higher-number IDs.

Figure 4.1 Seating Arrangement for the Experiment



Note: One square represents one desk. Students were seated in order of their ID numbers. Each number indicates the order of student ID numbers in each block. The student with the lowest ID number sat in desk 1, the one with the highest ID in desk 50. Rows 1-3 had yellow sheets on the desks with instructions not to write their name, thus creating the Blind group. Rows 4-10 had no extra sheets. In these rows students were expected to write their name to create the Visible group.

4.3.2 The Experiment and Data Collection

The students in each session arrive at the exam hall and locate their assigned block based on the course they are taking. Within the block they then locate their assigned

desk, which is marked with their student ID number. Once the exam session starts students have three hours to complete their exams. During that time one invigilator (not the same person as the exam grader) supervises each block. We asked the invigilators to place yellow sheets on all desks in the first three rows of each block (see Figure 4.1), thus ensuring that the recipients were mixed by ID number, and thus were more or less randomly treated by seniority in the University. The sheets stated that the students on whose desks one was placed should *not* write their name but *only* their ID number on the exam sheets (see Figure 4.6 in the appendix).[35] Because of the predetermined arrangement of desks this meant that a random sample of students within each course - the "*blind"* group - was asked not to write their names, so that the grader would only observe their ID numbers when grading. For the rest of the students - the *"visible"* group - graders could observe both names and IDs, as in previous teaching periods.

We collected additional information from several other sources. The Exams Office provided us with the nationality and gender of the students, grades in previous courses, and the desk arrangement during the exam. From the seating arrangement we could infer which students were asked not to write their names (yellow sheets, rows 1-3) and which were allowed to do so. To check students' compliance with the experiment's instructions, we manually went through all the exams and noted which students wrote down their names and which students did not.[36]

At the SBE it is common practice to split the grading burden among various graders by letting each one handle all the answers to a particular set of questions on the same exam. The course coordinators identified the grader of each question and provided us with information on the grading. This information included the score on each question and the maximum possible points per question. They also provided other grades that the student had attained in the course, including on course participation, presentation and any term paper.[37] A survey sent after the grading to all graders and course coordinators provided information on the grader's gender, nationality,

35. We placed the sheets on entire rows instead of scattered seats within each block for simplicity. We treated rows instead of columns in order to capture students with a variety of high and low ID numbers within each course. The Exams Office informed the course coordinators—who were in charge of organizing the grading of the exams—before the examination period that a new examination procedure was being tested, so that some exams might only have ID numbers. They were asked to have those exams graded as they usually would.
36. This was done immediately after the exam, before the course coordinators received the exams and started the grading process.
37. Most course coordinators had this information readily available in an Excel file. We manually collected the scores on each exam question for 7 courses.

teaching experience and grading behavior during the experiment.[38] From the SBE's online tool for course evaluations we gathered the total number of courses in which the grader had been involved at the SBE and the average instructor evaluations provided by students for that grader in all previous courses since the creation of the online tool. Our sample contains 25 out of the 42 courses that had final exams, including 42 different graders and 1,495 exams.[39]

The upper part of Table 4.2 examines the internal validity of the experiment, testing whether the questions in the treated (Visible) group were answered by students whose characteristics before they entered the examination room differed in measurable dimensions from those in the untreated (Blind) group. We present these results separately for those students whom we intended to treat (ITT) and those who were actually treated.[40] We first examine differences by gender and nationality, the two characteristics on which we focus, and in the students' grades before the final exam. The Blind and Visible groups are balanced in both gender and nationality: The p-values indicate that none of the tests of differences in the means between the Blind and Visible groups along the dimensions that form the focus of this study can reject the hypothesis that they are zero. Indeed, not only are the fractions of men and women, Germans and Dutch, insignificantly different from each other; the absolute differences between the Blind and Visible groups are never greater than two in the second decimal place.

We have additional information on some of the students - other grades that were received before the exams were given, such as prior grade point average (GPA), and classroom participation, presentation in class and term-paper grades in the particular course. We find no significant differences between the Blind and Visible students in GPA and their participation grades. The Visible group performs slightly better in the grades assigned for student presentations. This difference is not quite statistically significant, however; and perhaps more important, grades for classroom presentations were given to only about one-third of the students.

38. We manually added the gender and nationality of the graders who did not fill out the survey. Grading behavior includes whether graders looked up any names while grading.
39. We excluded 8 courses that only used Multiple Choice or Fill-In-The-Blank questions. In 7 out of the 34 eligible courses the coordinators either declined permission to use the data or did not respond to repeated requests for this information. We excluded one course for which the answer sheets did not ask for the students' names but only for their IDs and another course which did not hold the exam in the conference hall.
40. The blind treatment group had a little over 80-percent effectiveness, and an additional 2 percent of the students got into the blind group but should not have. This latter was most likely due to mistakes by the invigilators when placing the yellow sheets or by students forgetting to write their names.

Table 4.2 Student Characteristics by Intended and Actual Treatment Status

Internal validity: Pre-experiment		(1) Blind			(2) Visible			p-value of difference
		Mean	SD	N	Mean	SD	N	Blind-Visible
Female	ITT	0.369	0.483	452	0.352	0.478	1,043	[0.502]
	Treatment	0.363	0.482	399	0.355	0.479	1,096	[0.758]
German	ITT	0.374	0.484	452	0.353	0.478	1,043	[0.420]
	Treatment	0.373	0.484	399	0.354	0.478	1,096	[0.486]
Dutch	ITT	0.363	0.481	452	0.343	0.475	1,043	[0.452]
	Treatment	0.351	0.478	399	0.349	0.477	1,096	[0.932]
GPA	ITT	7.197	0.628	443	7.215	0.665	1,021	[0.607]
	Treatment	7.178	0.618	389	7.221	0.667	1,075	[0.241]
Participation	ITT	7.690	0.986	306	7.633	1.031	706	[0.386]
	Treatment	7.612	0.968	263	7.664	1.035	749	[0.452]
Presentation	ITT	7.795	1.164	191	7.930	1.059	436	[0.179]
	Treatment	7.758	1.172	181	7.942	1.055	446	[0.070]
Term paper	ITT	7.870	0.665	109	7.743	0.898	281	[0.126]
	Treatment	7.870	0.697	97	7.748	0.882	293	[0.166]
Internal validity: Within-experiment								
Multiple Choice	ITT	5.829	1.972	277	6.043	1.942	661	[0.128]
	Treatment	5.792	2.009	253	6.049	1.928	685	[0.078]
Fill-In-The-Blank	ITT	5.325	2.208	152	5.555	1.996	382	[0.264]
	Treatment	5.367	2.167	148	5.536	2.016	386	[0.411]

Note: The pre-experiment validity only includes students in the estimation sample. The within-experiment validity uses information on students who participated in the experiment, but the information on these answers is not part of our analysis. The p-values of differences between the Visible and Blind groups are calculated with clustered standard errors by student.

We also have grades from Multiple Choice and Fill-In-The-Blank questions that were included in a minority of the final exams. We can thus test whether, despite the apparent randomness of assignment, outcomes differed between the two groups on questions on which the grading was unambiguous and could not have been affected by the mechanisms we study here. As the bottom part of Table 4.2 shows, the Blind group did have marginally higher scores on the Multiple Choice questions, but here

too the differences are not quite statistically significant. These results confirm that the research design created equivalent groups of students.[41]

4.4 Inferring Average Outcomes and Distributions of Preferences

Let a student, denoted by *s*, answer an exam with several questions, and let the grader of each question be denoted by *g*. We index each answer by the pair (*s*, *g*).[42] We also know the pair (C(*s*),C(*g*)), where C is either some student-invariant bivariate characteristic, such as gender, or some characteristic vector, such as nationality. Finally, we know whether a particular answer by a particular student was graded blind or visible, so that each pair (C(*s*),C(*g*)) can be expanded to the triplet (C(*s*),C(*g*),*v*), where *v*=1 if the grading is visible and 0 if not.[43]

Consider the score function $S(C(s),C(g),v)$ for each exam question, where we are especially interested in examining how *S* varies between cases when *s* and *g* match (i.e. share a common characteristic) and when they do not, and how that variation is affected by *v*. Define the following indicators:

(1a) $I1\{(C(s),C(g), v)\} = 1$, if $C(s)=C(g)$ and $v=1$, 0 if not

(1b) $I2\{(C(s),C(g),v)\} = 1$, if $C(s)=C(g)$ and $v=0$, 0 if not

(1c) $I3\{(C(s),C(g),v)\} = 1$, if $C(s)\neq C(g)$ and $v=1$, 0 if not

and

(1d) $I4\{(C(s),C(g),v)\} = 1$, if $C(s) \neq C(g)$ and $v=0$, 0 if not.

The average score of all students is:

(2) $T = \theta_1 S^*(I1) + \theta_2 S^*(I2) + \theta_3 S^*(I3) + [1- \theta_1 - \theta_2 - \theta_3] S^*(I4)$

41. Considering that we tested several separate characteristics, it is not unlikely that some of those tests will reject the null hypothesis at the 10 percent level purely by chance. If we correct the p-values for multiple testing (using the Bonferroni, Šidák, or Holm adjustments), we find no significant differences between Blind and Visible students in any of the characteristics, even at the 10 percent level of significance.
42. We ignore course identifiers for simplicity, since all graders except one were uniquely assigned to one course.
43. Presumably all particular (*s*, *g*) combinations are either blind or visible (although we investigate the extent of blindness in the blind grading in Section 4.6).

where the weights θ_i are the shares of answers graded under each regime, and the (*) denotes an average over those answers.[44] Because we created the neutral categories with blind grading, we can estimate the average treatment effect on students for whom C(s) = C(g) (i.e., grader and student "match" on characteristic C) as:

(3a) $e^* = [S^*(I1) - S^*(I2)]$

and the treatment of students for whom C(i) ≠ C(g) (who do not "match" on C) as:

(3b) $x^* = [S^*(I4) - S^*(I3)]$

If graders are endophilic and exophobic, e^*, $x^* > 0$. Identifying endophilia and exophobia as e^* and x^* relies on the assumption that graders are neutral towards blind exams. In Section 4.5 we present estimates of each of the effects as discussed here. We discuss the implications of alternative behavioral assumptions in Section 4.6.

From Equation (2) we can also recover the average "total" effect of the characteristic C(s) for a particular value, C(s) = C'. This is particularly important if we want to address the question of whether disclosing certain information (such as gender or nationality) affects an outcome, given a distribution of preferences and graders. Consider a variant of (2):

(4) $T_C = \eta_1 S^*(I1|C(s)=C') + \eta_2 S^*(I2|C(s)=C') + \eta_3 S^*(I3|C(s)=C') + [1- \eta_1- \eta_2 - \eta_3]S^*(I4|C(s)=C')$

where the weights η represent the shares of answers graded under each regime for all students with characteristic C(s) = C'. The total treatment effect of a particular characteristic C' being observable is the weighted average of the treatments when C(g) = C' and when C(g) ≠ C'. Thus:

(5) $M_{C'} = (\eta_1 + \eta_2)[S^*(I1|C(g)=C') - S^*(I2|C(g)=C')] - (1- \eta_1 - \eta_2)[S^*(I4|C(g)=C') - S^*(I3|C(g)=C')]$

Equation (5) shows that the average treatment effect of a characteristic will depend on two factors: 1) The degree of endophilia and exophobia (the two bracketed

44. While the same average would apply for a n-fold characteristic if we focus only on whether or not C(s)=C(g), we could analogously and generally calculate n^2 average treatment effects, one for each of the n aspects of the characteristic compared to itself and each other aspect.

expressions); and 2) The share of questions that are graded by graders with matching characteristics ($\eta_1 + \eta_2$) versus non-matching characteristics ($1- \eta_1 - \eta_2$).

We can also observe the behavior of individual graders toward the student groups as defined by C(*s*). Each grader scores answers written by many different students, some with characteristics that match hers, others with characteristics that do not match, some of whom are graded Blind, others graded Visible. Then for a grader *g* we can calculate her average treatment of students, T^g, in a manner analogous to the average effect in (2) and obtain a distribution over all graders. More interesting for our purposes, we can estimate each grader's preferences for students who do and do not match their characteristics as:

(6a) $e^g = S^{*g}(I1)-S^{*g}(I2)$

and

(6b) $x^g = S^{*g}(I4)-S^{*g}(I3)$

where $S^{*g}(Ij)$, j=1,2,3,4, is the average over all students whose exams are scored by grader *g* under each regime I*j*. Using these grader-specific average treatments, we can then obtain non-parametric estimates of the distributions of endophilia and exophobia, *f(e)* and *h(x)*, as discussed in Section 4.4. Thus, in addition to being able to distinguish the average extent of favoritism toward one's own group from the average extent of discrimination against other group(s), the data allow us to obtain complete distributions of agents' implicit preferences.

4.5 Empirical Strategy and Basic Results

To estimate the impacts of nationality and gender matches on the points that graders assigned to students' answers, and to infer the differences discussed above, we estimate the regression:

(7) $S = \beta_1 MATCH*VISIBLE + \beta_2 MATCH*BLIND + \beta_3 NON\text{-}MATCH*VISIBLE + \beta_4 NON\text{-}MATCH*BLIND + \gamma'Z + \varepsilon$,

where here *S* is a unit normal deviate calculated for each exam question, and the other variable names are self-explanatory. The matrix Z includes nationality or gender indicators for both students and graders, ε is a zero-mean error term and the

regression is estimated without a constant. From this equation the estimates of the average extent of endophilia and exophobia are:

(8a) $e^* = S^*(I1) - S^*(I2) = \beta_1 - \beta_2$,

and:

(8b) $x^* = S^*(I4) - S^*(I3) = \beta_4 - \beta_3$.

Thus the estimates of (7) provide direct analogs to the concepts we seek to measure. Note that these calculations mean that endophilia (exophobia) is indicated by a positive e^* (x^*).

One special benefit that we obtain from our setting is that we can be sure that the implied preferences on matching are not being driven by confounding factors like unobserved heterogeneity. In our experimental setting we are comparing arguably identical groups whose only difference—because the treatment was random—is that the graders observed the names of some but not of other students. The experiment allows us explicitly to compare e.g., Visible to Blind German students. This means that anything specifically German, such as writing style in English or particular calligraphic patterns, washes out in this comparison. This framework also makes it easy to expand Equation (7) to include interactions with some of the graders' measurable characteristics and thus to examine how e^* and x^* vary with them. We deal with these extensions in Section 4.6.

The first two columns of Table 4.3 present the estimated β and their standard errors for the basic equations describing matches/non-matches along the criteria of nationality and gender. Since the experimental design randomized by blocks of students within each course, we cluster the standard errors at the Intention-To-Treat and course (ITT-course) level, allowing for two clusters per course. We focus throughout on the estimates of e^* and x^* and their statistical significance.

Table 4.3 Basic Estimates of the Extent of Favoritism and Discrimination by Nationality and Gender (N = 9,330)

	(1)	(2)	(3)			(4)	
	Nationality	Gender	Nationality			Gender	
Interaction with:	-	-	*German*	*Dutch*	*Other*	*Female*	*Male*
(1) MATCH*VISIBLE	0.287	-0.039	0.306	-0.012	-	0.156	-0.039
	(0.038)	(0.025)	(0.021)	(0.099)	-	(0.028)	(0.027)
(2) MATCH*BLIND	0.115	-0.099	0.165	-0.204	-	0.101	-0.101
	(0.081)	(0.039)	(0.101)	(0.106)	-	(0.075)	(0.042)
(3) NON-MATCH*VISIBLE	0.177	-0.076	0.148	-0.048	-0.123	0.150	-0.101
	(0.050)	(0.040)	(0.070)	(0.053)	(0.067)	(0.047)	(0.046)
(4) NON-MATCH*BLIND	0.172	-0.101	0.060	-0.095	-0.035	0.053	-0.071
	(0.057)	(0.047)	(0.080)	(0.077)	(0.072)	(0.038)	(0.079)
Endophilia [(1)-(2)]	0.172	0.060	0.140	0.193	-	0.055	0.062
p =	[0.028]	[0.140]	[0.171]	[0.049]	-	[0.471]	[0.188]
Exophobia [(4)-(3)]	-0.005	-0.025	-0.088	-0.047	0.088	-0.097	0.030
p =	[0.904]	[0.700]	[0.095]	[0.528]	[0.174]	[0.124]	[0.740]
Adj. R2	0.015	0.009	0.016			0.010	

Note: Standard errors in parentheses and p-values in square brackets. Both are clustered by ITT-Course. Columns (1) and (2) present the estimates of Equation (7) without a constant. Columns (3) and (4) are based on Equation (7), with the main variables interacted with CHARACTERISTIC, where CHARACTERISTIC are indicators for nationality in (3) and for gender in (4). MATCH*Other interactions in (3) are empty because we define MATCH = 1 only for German and Dutch students. Other nationalities almost never matched. Main effects are included throughout, when not perfectly collinear with the main coefficients.

It is clear that there is substantial endophilia by nationality in the grading. A student who matches the grader's nationality receives a score that is 0.17 standard deviations higher when her name is visible than when it is not. This addition to a matched student's grade is statistically significant at conventional levels. This effect is also economically important: Given that all the scores have been unit-normalized, this effect is equivalent to moving from the median score to the 57th percentile of the distribution of scores. Its magnitude is similar to that of the effect of large differences in teacher quality on students' test scores that was found by Rivkin *et al* (2005). While favoritism by nationality exists in grading, there is no apparent exophobia by nationality: The estimated impact of being visible when not matching by nationality is small and positive.

The results of estimating the regression examining gender matching are shown in Column (2) of Table 4.3. Although the point estimate suggests the existence of a small degree of endophilia, we cannot reject the hypothesis that it is zero. For non-matches there is exophilia, but here too the impact is statistically insignificant and also minute. On average grading seems gender-neutral in all dimensions.[45]

Going behind the information in Columns (1) and (2), we can ask whether, for examples, endophilia by nationality is the same for Dutch and German graders, and whether the absence of endophilia or exophobia exists for both male and female graders. We do this by expanding Equation (7) to include interactions of student nationality or gender with *MATCH*VISIBLE*, *MATCH*BLIND*, *NON-MATCH*VISIBLE*, and *NON-MATCH*BLIND*. Columns (3) of Table 4.3 show the estimates of this expanded specification by nationality. A comparison of the results suggests that endophilia by nationality arises more from the behavior of Dutch than of German graders, although the difference between the two point estimates is not statistically significant.

Columns (4) of Table 4.3 show estimates of expanding Equation (7) by gender. The results look very much like those in Column (2): Neither male nor female graders exhibit significant endophilia or exophobia, and for both men and women the absolute impacts are small. Again, there is no sign of either statistically significant or important differences in behavior depending on the match or non-match of the grader's and student's gender.

4.6 Robustness and Extensions

4.6.1 Treatment Failures

In interpreting these main results it is important to note that there are two potential sources of slippage in our treatment: Some students did not comply with the experimental instructions shown in the appendix and mistakenly wrote their names on the exam sheets; and some graders may have looked up at least some of the students' names.[46] To account for the first source of slippage we re-estimated the models described in the first two columns of Table 4.3 using intention to treat (ITT)

45. The results are also essentially the same when we include additional controls for seat number (see Figure 4.1) and the student's prior GPA.
46. Evidence on the magnitude of the first type of slippage can be seen in Table 4.2 in the differences between ITT and Treatment.

as an instrument for *VISIBLE*. As the first two columns of Table 4.4 show, the results are qualitatively identical to the ones in Table 4.3.

To account for the second source of slippage - that the grader may have been able to identify the characteristic of the Blind group - in the post-grading survey we asked graders whether they looked up any names on the exams that only contained ID numbers. Six of the thirty-three graders who responded to the survey acknowledged having done this. When we re-estimated (7) including only those graders who explicitly stated that they did not look up names, the estimated endophilia by nationality is the same and is even more precisely estimated. The last column of Table 4.4 shows there is no significant endophilia but significant exophilia by gender among those graders who did not look up names. The results of both slippages suggest that, if anything, our results understate the true extent of favoritism by nationality and gender.

Table 4.4 The Effects of Treatment Slippage by Students and Graders on Estimates of Endophilia and Exophobia

	(1)	(2)	(3)	(4)
	Nationality	Gender	Nationality	Gender
Regression:	IV		Did Not Look Up Names	
Endophilia	0.193	0.090	0.174	0.009
p =	[0.034]	[0.190]	[0.009]	[0.856]
Exophobia	-0.033	-0.039	-0.008	-0.119
p =	[0.538]	[0.595]	[0.878]	[0.010]
N	9,330	9,330	5,108	5,108
Adj. R2	0.015	-0.001	0.015	0.007

Note: p-values in squared brackets, based on standard errors clustered at the ITT-Course level. We report linear combinations based on extensions of Equation (7). Columns (1) and (2) are based on an instrumental variable regression (IV) estimated by 2SLS, where we use the intention to treat (ITT) to instrument for the treatment. The F-tests for the instruments always strongly reject the null. Columns (3) and (4) are based on Equation (7) using only graders who did not look up any of the names in the Blind group of exams. Main effects are included throughout.

4.6.2 Alternative Behavioral Assumptions

So far we have implicitly assumed that the graders are indifferent toward "blind" exams and treat these groups as a neutral baseline against which we measure endophilia and exophobia. Graders, however, might also form rational expectations about the "blind" exams, considering the underlying *distribution* of characteristics of students who wrote those exams and might score them accordingly. Let A_g be the share of students in the course who match grader g on the characteristic of interest, and let e^{re} and x^{re} be the grader's latent endophilic and exophobic preferences. Under rational expectations we can rewrite Equation (7) as:

(7') $S = e^{re}\cdot\text{MATCH*VISIBLE} + x^{re}\cdot\text{NON-MATCH*VISIBLE} - [e^{re}\cdot A_g + x^{re}\cdot(1-A_g)] + \gamma_2'Z + \varepsilon,$

where, from the grader's perspective, the students can either visibly match him, visibly not match him, or be in the Blind group (the omitted category). Equation (7') specifies that the grader will treat the students in the Blind group as the weighted average of how he would have treated students who matched him or not, with weights based on the characteristics of the Visible groups.

To determine whether assuming rational expectations about the Blind group's students can alter our results, we estimate (7') by non-linear least squares. The results confirm our main findings: We again find endophilia by nationality, although of slightly lesser but still statistically significant magnitude (0.133 standard deviations, p=0.025). We find no endophilia by gender and no exophobia by either gender or nationality. Moreover, the root mean square error of estimating (7') exceeds that of the estimate of (7). Because the blind-as-neutral assumption fits the data better, we continue defining endophilia and exophobia as discussed in Section 4.4 throughout the rest of the chapter.

4.6.3 Prior Grader-Student Contact, and Exam Type

The graders and exams differ along several dimensions on which we have information and which might affect their ability or interest in favoring/discriminating for/against students. We first look at whether the graders knew the students they graded, and thus whether endophilia/exophobia is present towards anonymous and familiar students alike.[47] We have no specific hypothesis on

47. The assignment of students and teachers to classes within a course is done by the Scheduling Department of the SBE, which does not consider students' preferences for particular teacher or

this possibility. On the one hand, it could be that prejudices are overridden by personal experience with the students. If so, discriminatory preferences will be stronger toward unknown students. On the other hand, it might not be the characteristic *per se* that the graders pay attention to, but something that graders can only observe on students with whom they interact. In this case discriminatory preferences will be stronger toward and against students whom the grader knows.

We construct an indicator of whether the grader may know a student based on whether the grader also taught him or her. Most of the teaching at the SBE is done in tutorials of 10 to 15 students for about 10 sessions in each seven-week block, so teachers have a fair chance to get to know their students. Some graders taught none of the students they graded, others taught all of the students they graded. By this measure the median grader knew 47 percent of the students graded (although obviously in most cases the grader could not identify individual students in the Blind group).

The first two columns of Table 4.5 present re-estimates of Equation (7), expanded to include interactions of the Know indicator with the four Match/Visible variables. The results show that endophilia by nationality is only present when graders did not know the students. This effect is twice as large as the mean effect in the baseline model. There is no evidence of exophobia by nationality regardless of whether the grader knew the student or not. There is evidence of endophilia and exophilia by gender, but again only when the grader did not know the student.

The exams at the SBE differ in the extent to which they have mathematical questions, depending mostly on the nature of the courses. Answers on the more mathematical exams are arguably less ambiguous, so that showing favoritism/discrimination on them might be more difficult. To separate the more from the less mathematical exams we asked three raters (from the SBE's pool of potential graders) to rate the exams as mathematical or not. Two of the three agreed in their categorizations of all the exams, while the third agreed in 80 percent of the cases. We thus created an indicator for Mathematical when at least two of the three raters designated an exam as such, which occurred for 9 out of 25 exams.

The third and fourth columns of Table 4.5 present estimates of Equation (7), expanded to include interactions of the Mathematical indicator with the main

---

teachers' preferences for a particular class. (See Chapter 5 for a detailed explanation on the assignment of students and teachers to classes at the SBE.) Also, the students have no way of knowing *ex ante* who their grader will be.

variables. The point estimates suggest that endophilia by nationality is stronger for less mathematical exams. The point estimates for exophilia by nationality and endophilia by gender are also significant for the more mathematical exams. This latter result is surprising, as one might expect that Blind exams might be less likely to be assignable to nationality or gender based on handwriting styles if the exam is more mathematical. None of the other results in the two columns is statistically significant.

Table 4.5 Endophilia and Exophobia When Graders Know the Students They Grade, and When the Exams are Mathematical (N = 9,330)

		(1) Nationality	(2) Gender			(3) Nationality	(4) Gender
Grader knows the student?:				*Exam was mathematical?:*			
Endophilia	*No*	0.320	0.120	Endophilia	*No*	0.228	0.034
	p =	[0.003]	[0.042]		p =	[0.039]	[0.578]
	Yes	0.052	-0.001		*Yes*	0.060	0.094
	p =	[0.580]	[0.980]		p =	[0.375]	[0.042]
Exophobia	*No*	-0.070	-0.112	Exophobia	*No*	0.055	0.002
	p =	[0.120]	[0.040]		p =	[0.345]	[0.975]
	Yes	0.070	0.085		*Yes*	-0.095	-0.081
	p =	[0.427]	[0.396]		p =	[0.020]	[0.197]
F-test differences:		[0.066]	[0.166]			[0.024]	[0.584]

Note: p-values in squared brackets, based on standard errors clustered at the ITT-Course level. We report linear combinations based on extensions of Equation (7). Columns (1) and (2) report interactions of the main variables with GRADERKNOWSSTUDENT, Columns (3) and (4) of the main variables with MATHEMATICALEXAM. F-test differences reports the p-values from testing the null hypothesis that Endophilia and Exophobia are equal for the groups defined by GRADERKNOWSSTUDENT and MATHEMATICALEXAM, respectively. Main effects are included throughout.

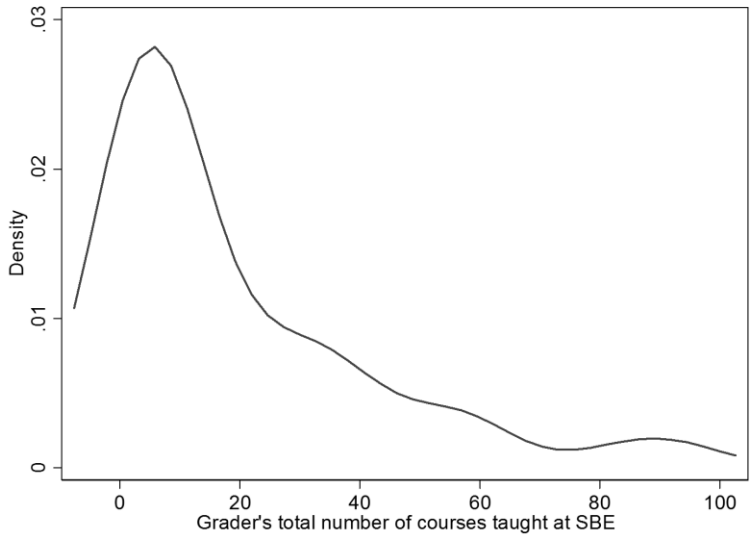
4.6.4 Distinguishing by Graders' Other Characteristics

We also examine whether discrimination or favoritism varies with grader experience and grader quality. We measure grader experience at this University as the number

of separate courses taught or tutored during the grader's tenure. We have no hypotheses about how university-specific experience might mitigate or exacerbate endophilia/exophobia. On the one hand, the set of more experienced graders may exclude those whose behavior was so egregiously unfair that the University did not renew their contracts. On the other hand, more experienced graders may be secure in their positions and feel able to indulge their preferences for students who match their characteristics and/or against those who do not.

The total number of courses taught/tutored at the University since the online data became available (including the courses we are using here) ranges from 1 to 94; the 5th, 50th and 95th percentiles, for which we present estimation results, are 1, 8 and 59 courses.[48] Figure 4.2 shows the kernel density of courses taught by grader, which demonstrates the distribution's very long right tail. The first and second columns of Table 4.6 present re-estimates of Equation (7), expanded to include interactions of grader experience with the four match/visible variables.

Figure 4.2 Kernel Density of the Distribution of Grader Experience



While the point estimate of the extent of endophilia by nationality is almost identical to the estimate in Table 4.3 at the median value of grader experience, it is not quite significantly nonzero. Rather, the significant average endophilia shown in Table 4.3

48. 59 and 94 might seem outlandishly large; but at this University there are 6 teaching blocks in each academic year, so it is not difficult to accumulate 50 or more courses of experience.

results disproportionately from the behavior of the more experienced graders. By inference, they feel less inhibited about indulging their preferences for students who match their nationality. Inexperienced graders, perhaps because they feel themselves to be under greater scrutiny, show no significant endophilia (although the point estimate of their behavior is 60 percent of that of highly experienced graders). As with the basic estimates, there is no evidence of exophobia by nationality at any level of grader experience. The results by gender remain very similar: Just as at the sample means, so too at various levels of grader experience the parameter estimates show no sign of any significant endophilia or exophobia. The exception is the evidence of exophilia by gender for the most experienced graders.

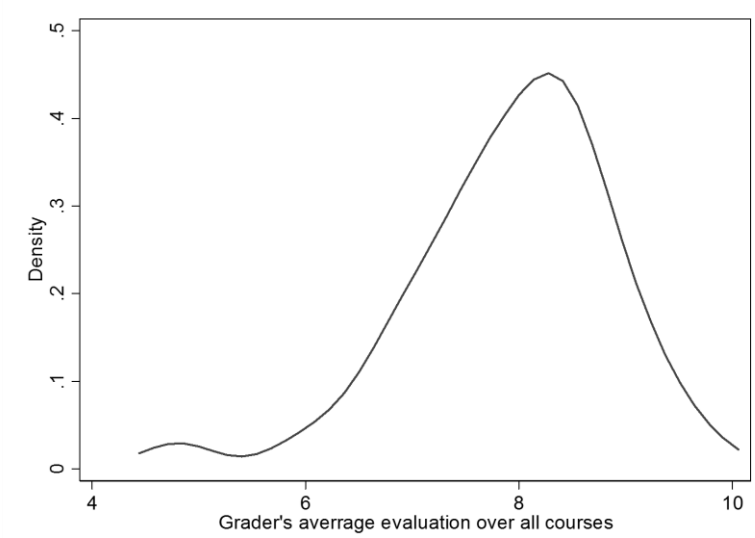
Table 4.6 Effects of Grader Experience and Grader Teaching Quality on Outcomes (N = 9197)*

		(1)	(2)	(3)	(4)
	Percentile:	Nationality	Gender	Nationality	Gender
At the mth percentile of:		*Experience*		*Teacher Quality*	
Endophilia	5th	0.154	0.076	0.307	0.039
	p =	[0.162]	[0.141]	[0.020]	[0.575]
	50th	0.166	0.073	0.170	0.068
	p =	[0.097]	[0.118]	[0.130]	[0.108]
	95th	0.248	0.045	0.048	0.093
	p =	[0.001]	[0.517]	[0.651]	[0.158]
Exophobia	5th	-0.024	0.008	-0.014	-0.140
	p =	[0.639]	[0.919]	[0.859]	[0.048]
	50th	-0.016	-0.007	-0.005	-0.013
	p =	[0.718]	[0.920]	[0.896]	[0.842]
	95th	0.045	-0.121	0.002	0.100
	p =	[0.635]	[0.053]	[0.962]	[0.289]
F-test interactions:		[0.547]	[0.261]	[0.356]	[0.075]

*p-values in square brackets, based on standard errors clustered at the ITT-Course level. We report linear combinations based on extensions of Equation (7). Columns (1) and (2) interact the main variables with TEACHEREXPERIENCE and evaluate the linear combinations at different percentiles. Columns (3) and (4) do the same with TEACHERQUALITY. F-test interactions reports the p-values from testing the joint significance of interactions of Endophilia and Exophobia with TEACHEREXPERIENCE and TEACHERQUALITY, respectively. Main effects are included throughout.

We measure grader quality as the average of all the evaluations that the instructor received from students during her career at the University. Evaluations are given on a ten-point scale. In our sample the averages range from 6.5 to 9.2, with the 5th percentile being 7.1, the median being 8.0, and the 95th percentile equaling 8.8. As Figure 4.3 shows, the distribution of average evaluations is quite close to symmetric.

Figure 4.3 Kernel Density of the Distribution of Student Evaluations of Graders



We interact the grader's average instructional evaluation with all the variables in Equation (7) and present the results in Columns (3) and (4) of Table 4.6. Our finding of endophilia by nationality at the mean demonstrated in Table 4.3 arose from behavior that varies sharply with the regard in which graders have been held by students. Those graders/instructors who have been rated highest by students show no significant endophilia, and the point estimate of this effect is small. An instructor whose teaching has been rated at the median of this measure behaves much like the mean instructor—substantially favoring those who match her nationality, unsurprisingly given the symmetry in the distribution of teaching evaluations. The worst-rated instructors, however, favor those students who match their nationality much more strongly than does the median or average instructor. Implicitly a poorly rated instructor raises the score of the median student who matches her nationality from the mean to the 61st percentile of the distribution of scores. There is no evidence of exophobia by nationality. In a similar fashion, the little evidence there

was of exophilia by gender seems to be driven by the worst-rated teachers. In sum, worse teachers behave differently from better ones, favoring students of their own nationality and, to a lesser extent, the other gender.

4.7 The Average Treatment Effect of Visible Student Characteristics

To evaluate whether the visibility of names differentially favors or disadvantages certain groups of students, and also to see how these students would be affected by the introduction of anonymous grading, we calculate the average treatment effect (ATE) of each characteristic's visibility. Recall from Equation (5) that the ATE can be calculated as the difference between endophilia and exophobia, each weighted by the share of questions that was graded by graders with matching and non-matching characteristics. Table 4.7 shows the ATE of being seen as German, Dutch, or any other nationality, and of being seen as female or male. The point estimates for German and Dutch students are similar in size and (marginally) significantly positive, demonstrating that both German and Dutch students benefit from visible grading. The point estimates further suggest that other nationalities are disadvantaged by it, although the ATE is not statistically significant. Even if they are not suffering from an absolute disadvantage, however, the notion that other nationalities are disadvantaged becomes straightforward for situations in which they compete with German and Dutch students. An example is the allocation of student exchange positions at popular universities abroad, which is done based on relative grades. The difference between Germans and others is significant ($p=0.004$), as is the difference between Dutch and others ($p= 0.025$). Consistent with our previous results, the point estimates for females and males are positive but smaller in size.

Columns (1) to (4) of Table 4.7 decompose the ATE by showing endophilia and exophobia (Columns (2) and (4)) and the share of students with the given characteristic that was graded under each regime (Columns (1) and (3)). (The estimated effects of endophilia and exophobia are taken from Table 4.3.) The ATE for German and Dutch students is small because of the relatively small shares of questions that are graded by graders of the same nationality. It is easy to simulate the sizes of these effects for a situation in which a large share of the students in either category were matched to the graders. Notice also that the mix of graders is not always the most important determinant of the ATE: The difference between the

effects when matched and not matched for females is rather small, so that the ATE will be small regardless of the gender mix of graders.

Table 4.7 The Average Treatment Effect (ATE) of the Visibility of Student Characteristics

Student	Total ATE	p-value	(1) Share matched *(η1+η2)*	(2) Endophilia	(3) Share not matched *(1-η1-η2)*	(4) Exophobia
German	0.103	[0.049]	0.29	0.140	0.71	-0.088
Dutch	0.107	[0.067]	0.41	0.193	0.59	-0.047
Other	-0.088	[0.174]	-	-	1	0.088
Female	0.078	[0.142]	0.45	0.055	0.55	-0.097
Male	0.027	[0.548]	0.62	0.062	0.38	0.030

Note: The ATE is calculated as shown in Equation (5). The p-values are based on standard errors clustered at the ITT-Course level. Columns (1) and (3) show the share of questions, for a given characteristic, which were graded by graders with matching and non-matching characteristics. Columns (2) and (4) show the ATE on the treated, as reported in Table 4.3.

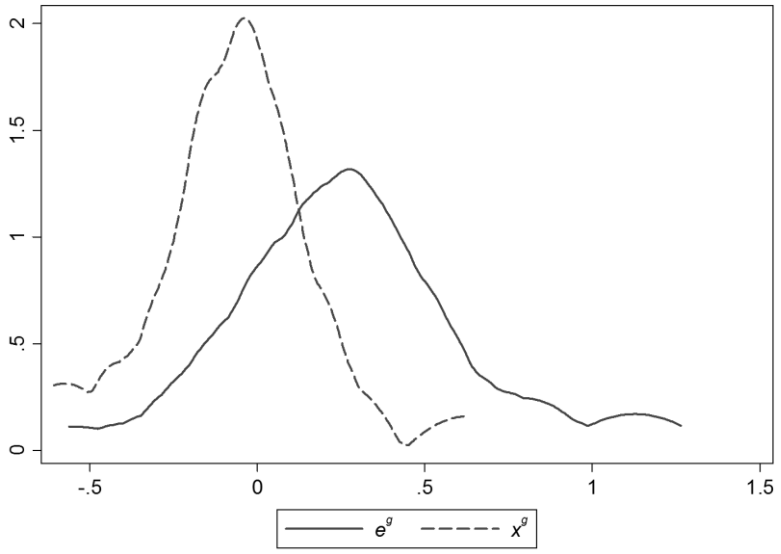
4.8 Heterogeneity in the Distributions of Preferences

The results thus far describe either average responses of endophilia or exophobia by nationality or gender over all graders, or examine how this behavior differs in relation to a few of the graders' specific characteristics. This parallels but expands upon the focus in the literature on average differences between groups. In this section we move to a different dimension, the distribution of implicit tastes for favoritism and discrimination, first considering the shapes of the entire distributions of graders' preferences and then calculating their correlations, as suggested by the theoretical discussion.

To obtain a feel for why examining heterogeneity in preferences might be interesting, consider the kernel density estimates of the graders' endophilia and exophobia by nationality, shown in Figure 4.4, and their kernel densities by gender,

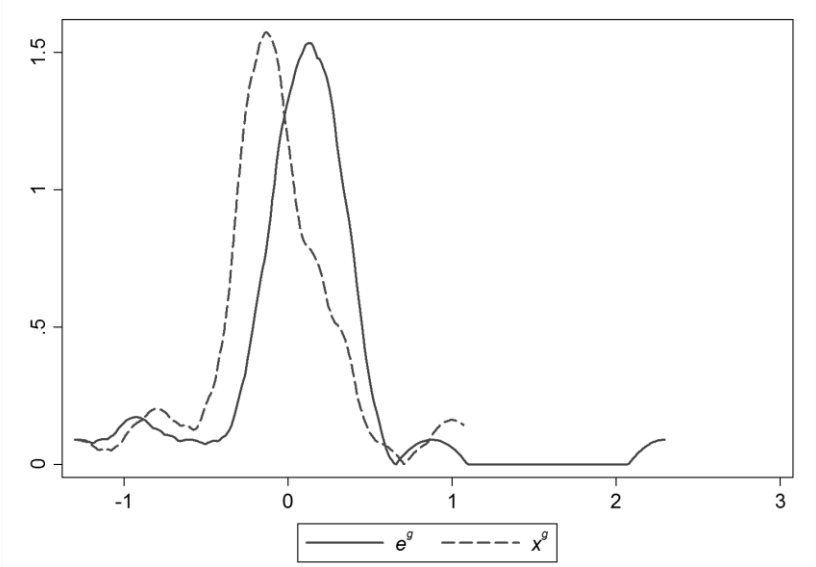
shown in Figure 4.5. Each kernel is based on those graders for whom we could infer the extent of both endophilia and exophobia (for nationality, 24 graders, for gender, 38 graders).[49] The estimates along the criterion of nationality suggest that preferences are distributed fairly symmetrically, in the case of endophilia around a positive mean, and around zero in the case of exophobia. Both densities are consistent with our inferences in Table 4.3 about the mean effects. A similar conclusion is suggested by the kernels of endophilia and exophobia by gender, although there are a few outliers.[50]

Figure 4.4 Kernel Density Estimates of Graders' Preferences by Nationality



49. We derive the shape of the graders' preferences based on the estimates of e^g and x^g calculated as in equations (6a) and (6b). We infer these two measures for each grader based on how each scores the student who does or does not match them under the blind and visible regimes.
50. We can examine whether extreme values in the distributions of preferences for nationality or gender are driving our mean effects. We trim those graders with the most extreme preferences from the samples, dropping the two most extremely endophilic/endophobic and exophobic/exophilic graders in each case. Despite the small amounts of asymmetry in some of the distributions, trimming does not qualitatively alter the conclusions about the absence of endophilia or exophobia by gender on average, nor does it alter the conclusions about these outcomes by nationality.

Figure 4.5 Kernel Density Estimates of Graders' Preferences by Gender



By observing the entire distribution of preferences we can also test two hypotheses: 1) There is evidence of endophilia or exophobia in the overall distribution (not just at the mean), and 2) There is heterogeneity in endophilia or exophobia among graders. Testing these two hypotheses is equivalent to testing whether $e^g=0$ ($x^g=0$) for all g, and whether the e^g (x^g) are equal to each other for all g, respectively. The F-tests of these hypotheses (eight in total) all reject the null hypothesis at conventional significance levels, showing that endophilia and exophobia in both nationality and gender are real phenomena (even though at the means only endophilia by nationality seems to matter), and that there is significant heterogeneity in these preferences across graders.[51]

As we showed in Section 4.2, the impact of the interaction of endophilia and exophobia depends on their correlation across potentially discriminating agents. In our data the correlations are -0.36 for preferences on nationality, and -0.16 for preferences on gender (weighting each grader by the number of students graded). Those who are more endophilic are less exophobic. Interestingly, and remarkably, in the GSS data summarized in Table 4.1, the correlations are in the same direction:

51. The demonstrated heterogeneity of preferences should reduce any concern about the absence of exophobia at the mean because the possibly large (psychological) costs of giving lower grades. Also, it should vitiate concerns about our behavioral assumption on how graders treat Blind exams, since it shows that Visible non-matches are treated differently from Blind ones by most graders.

Those Whites who feel closer to Whites also feel closer to Blacks, and to roughly the same extent as implied by behavior in our sample.

4.9 Conclusions and Implications

We have demonstrated that what is called discrimination—a relative difference in outcomes between two groups—is composed of differential treatment of the in-group and the out-group, and that it is possible in real-world situations to measure the sizes of these two components simultaneously. In our example we find that most of the apparent discrimination by nationality results from substantial endophilia and that there is no evidence on average of exophobia. We find some evidence of graders favoring the opposite gender on average, though it is less definitive.

These are average effects. At least as interesting is the heterogeneity in the demonstrated preferences of the individuals deciding how to treat those who match or do not match them. We have further shown that apparently discriminatory outcomes can be vitiated in a variety of ways, operating both on the endophilic and exophobic preferences of the discriminating agents and the share of matching and non-matching characteristics.

We also show the importance of measuring the relation between endophilia and exophobia in the labor market: Their joint distribution will influence market-based measures of discrimination. This result makes it even clearer that they are non-redundant measures. It also forces us to reconsider what we know about the effectiveness of anti-discrimination policies and the advances against discrimination in the labor market. The change over time in measures of discrimination, such as the market discrimination coefficient, may not only reflect a change in the means and variances of the distributions of expressed preferences. It may also reflect a change in the correlation between endophilia and exophobia. This changing correlation might explain the unchanging Black-White wage gap over a period when racial attitudes appear to have become more tolerant.

Assuming that the dominance of endophilia over exophobia that we have demonstrated for nationality is ubiquitous in labor markets, the fact has important implications for the measurement of "discrimination" in labor markets. Decompositions that adjust a gross wage differential into parts due to different characteristics or different treatments in the labor market can be made using either the majority or the minority wage as the base case. In the literature (e.g., Neumark, 1988; Booth *et al*, 2007; Elder *et al*, 2010) that discusses these decompositions of

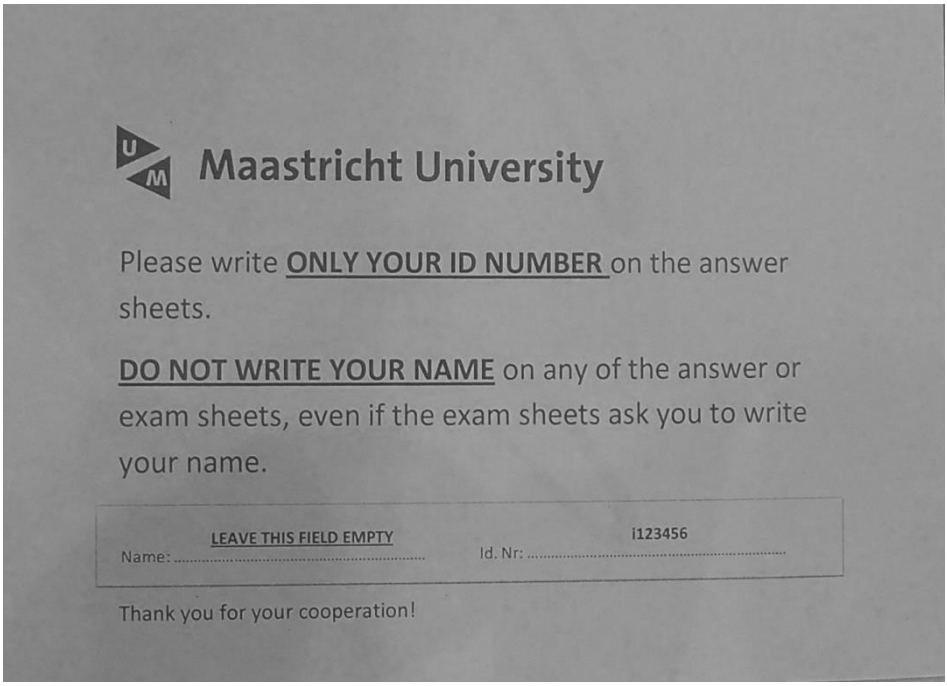
wage differentials (by race, gender, and many others) a crucial question has been which group's actual wage to treat as the baseline. Endophilia dominating exophobia would suggest using the minority group's wage as the baseline and adjusting the wages of the majority. More generally, if we knew the relative importance of each type of behavior, the appropriate treatment would be a weighted average of the different methods of decomposition.

Having shown that we can distinguish endophilia from exophobia, it is also worth considering how policy might be tailored to reduce relative differences arising from prejudice. Assume that our results carry over to the labor and other markets, and that endophilia is the main source of apparently discriminatory outcomes. If so, we can infer, for example, that moral suasion that stresses to members of the majority group that minority-group members are not "bad" might be ineffective.

Can the distinctions that we have defined and measured here be inferred in the still more important labor-market context using actual wage and/or employment outcomes? One might imagine cases where a majority group deals with several minority groups, about one of which it feels demonstrably neutral. In that case too endophilia and exophobia (toward the other minorities) are identifiable. So too, one might link differences in economic outcomes to information on attitudes in a population about one's own and other groups. The main point is that these preferences generate different outcomes with different distributions of welfare, so that determining their relative sizes is economically important and, as we have shown, possible.

4.10 Appendix

Figure 4.6 Yellow Sheet Placed on Some Students' Desks Before the Exam.



Maastricht University

Please write **ONLY YOUR ID NUMBER** on the answer sheets.

DO NOT WRITE YOUR NAME on any of the answer or exam sheets, even if the exam sheets ask you to write your name.

LEAVE THIS FIELD EMPTY	i123456
Name:	Id. Nr: .....

Thank you for your cooperation!

Chapter 5
On the Nature of Peer Effects in Academic Achievement*

* This chapter is based on joined work with Ulf Zölitz. We thank Joël Castermans, Sanne Klasen and Kim Schippers from the SBE Scheduling Department and Sylvie Kersten from the SBE Exams Office and Jeannette Hommes from the Educational Research and Development Department for providing data and valuable background information. We further like to thank Sandra Black, Lex Borghans, Andries de Grip, Thomas Dohmen, Armin Falk, David Figlio, Bart Golsteyn, Daniel Hamermesh, Nicolás Salamanca, Benedikt Vogt and participants at various seminars and conferences for helpful discussions and comments.

5.1 Introduction

Identifying peer effects is empirically challenging because the reason why a student is in a certain peer group (school, grade, section, dorm, etc.) is likely to be correlated with factors unobserved to the researcher which also affect student performance. The most convincing peer effects studies have therefore exploited random assignment of students to peer groups. This, however, has only been done in very particular situations and/or for very particular peer groups. At the university level studies have exploited (conditionally) random assignment of students to dorm rooms (e.g. Brunello, De Paola, & Scoppa, 2010; Sacerdote, 2001; Zimmerman, 2003) and conditionally random assignment of students to living communities in military colleges (Carrell, Fullerton, & West, 2009; Carrell, Sacerdote, & West, 2013; Lyle, 2007). At the pre-university level, Duflo, Dupas and Kremer (2011) have randomly assigned students to classes in elementary schools in rural Kenya. It is not clear how findings from these contexts generalize to the western classroom. The school classroom or university section is the unit where most of the peer interaction takes place and where students are forced to interact with each other. The classroom/section is, as compared to the dorm room or living community, also the most policy relevant peer group because it can most easily be influenced.

In this chapter we exploit random assignment of students to sections to study peer effects at the university level. Our dataset consists of all students enrolled at the School of Business and Economics (SBE) at Maastricht University over a period of three years, which amounts to 7,746 students and 41,749 student grades. Course participants are assigned to sections, groups of 10 to 15 students, which spend most of their contact hours together in one classroom.[52] Our measure of student performance is course grades. Following the standard approach in the literature, we use a pre-treatment indicator of peer quality: the GPA of the peers, to avoid the reflection problem as described by Manski (1993).

In the linear-in-means specification, we find that being assigned to a section with on average higher ability peers increases students' grades in that course by a statistically significant, but small amount. One standard deviation increase in the average peer GPA causes a .0116 standard deviation increase in student grade. As expected under random assignment, the size of our estimate is unaffected by the inclusion of large sets of teacher fixed effects and fixed effects for other courses taken at the same time. The peer effects we identify are heterogeneous and point to

52. Throughout the chapter we will use the word "section" when we refer to the tutorial group to which the students get assigned within each course.

an inverse U-shape relationship between peer quality and own performance. High and medium ability students are positively affected by high ability peers and negatively affected by low ability peers. Low ability students seem to benefit from the presence of medium rather than high or low ability students. Taken together our findings suggest that students benefit from better peers as long as the difference between own and peer ability is not too large.

This work is to the best of our knowledge the first that uses random assignment to estimate peer effects at the university section level.[53] Our study closely relates to Carrell et al. (2013) who study peer effects at the squadron level at the United States Air Force Academy (USAFA).[54] The authors exploit random variation in peer quality and find that low ability students benefit from high ability peers. Based on these results they then conduct an experiment designed to increase performance of low ability students by assigning them to squadrons with a large fraction of high ability peers. This intervention, however, failed and had perverse effects on low ability students who were supposed to be helped by the intervention. Contrary to their own previous finding, but consistent with the evidence we present in this chapter they find that, low ability students were negatively affected by an increase in the share of high ability peers.[55]

The roommate literature, which exploits conditionally random assignment of students to dorms, generally finds small positive or no significant peer effects (Brunello et al., 2010; Sacerdote, 2001; Zimmerman, 2003). De Paola and Scoppa (2010) investigate peer effects in a sample of 212 Italian university students and use random first year assignment of students as an instrument for current university peers. They find that a one standard deviation increase in peer quality measured as first year GPA increases exam performance by .11 standard deviations (OLS estimation) and .19 (IV estimation). Martins and Walker (2006) use the fact that the

53. Peer effects at the university level, as compared to pre-university education, might not only be different in size but also follow different dynamics. On the one hand, university students, as opposed to students in pre-university education, are more homogeneous, spend less time in the classroom and invest a greater share of their time in self-study. This would suggest that peer effects are generally less important at the university level. On the other hand, students learn arguably more from each other in classroom discussions and study groups and comparably less from the teachers.
54. The squadron is the living community at the campus of the USAFA where students have much of their social and academic interactions. Sections, however, are composed of members of different squadrons.
55. Carrell et al. (2013) explain this finding with new patterns of social interactions caused by the new peer group assignment. In the same context, Carrell *et al.* (2009) find positive peer effects in linear-in-means specification at the squadron level. In a similar setting at the West Point Military College, Lyle (2007) does not find significant peer effects but this might be due to limited variation in peer quality.

allocation of students to sections conducted alphabetically by last names. In their sample of 441 students they do not find significant effects of section peers.

The remainder of the chapter is structured as follows: Section 5.2 provides information about the institutional environment, and the assignment procedure of students to sections. Section 5.3 discusses the dataset. Section 5.4 provides evidence that the assignment to sections is random, controlling for scheduling constraints. Section 5.5 describes our empirical strategy. Section 5.6 shows the estimation results. Section 5.7 concludes.

5.2 Background

5.2.1 Institutional Environment

The School of Business and Economics (SBE) of Maastricht University is located in Maastricht, a city in the south of the Netherlands. Currently there are about 4,200 students at the SBE enrolled in Bachelor, Master and PhD programs. Because of its proximity to Germany, it has a large German student population (53 percent) mixed with Dutch (33 percent) and other nationalities. About 37 percent of the students are female. The academic year at the SBE is divided into four regular teaching periods of two months and two skills periods of two weeks. Students usually take two courses at the same time in the regular periods and one course in the skills period. We exclude courses in skills periods from our analysis because these are often not graded and we could not always identify the relevant peer group.[56]

The courses are organized by course coordinators, mostly senior staff, and most of the teachers are PhD students and teaching assistants. Each course is divided into sections of maximum 16 students. These sections are the peer group we are focusing on. The course size ranges from 1 to 638 students and there are 1 to 43 sections per course. The sections usually meet in two weekly sessions of two hours each. Most courses also have lectures which are followed by all students of the course and are usually given by senior staff.

The SBE differs from other universities in its focus on Problem Based Learning (PBL).[57] The general PBL setup is that students generate questions about a topic at the end of one session and then try to answer these questions through self-study. In

56. In some skills courses, for example, students are scheduled in different sections but end up sitting together in the same room.
57. See http://www.umpblprep.nl/ for a more detailed explanation of PBL at Maastricht University.

the next session the findings are discussed with the other students of the section. In the basic form of PBL the teacher takes only a guiding role and most of the learning is done by the students independently. Courses, however, differ in the extent to which they give guidance and structure to the students. This depends on the nature of the subject covered, with more difficult subjects (e.g. Quantitative Methods) usually requiring more guidance, and the preference of the course coordinator and teacher.

Compared to the traditional lecture system, the PBL system is arguably more group focused because most of the teaching happens in small groups in which group discussions are the central part of the learning process. Much of the students' peer interaction happens with members of their section, either in the sessions, during work for common projects or in homework and study groups.

5.2.2 Assignment of Students to Sections

The Scheduling Department of the SBE assigns students to sections, teachers to sections, and sections to time slots. Before each period, there is a time frame in which students can register online for the courses they want to take. After the registration deadline, the scheduler gets a list of registered students and allocates the students to sections using a computer program. About ten percent of the slots in each group are initially left empty and are filled with students who register late.[58] This procedure balances the amount of late registration students over the sections. Before the start of the academic year 2010/11, the section assignment for Master courses and for Bachelor courses was done with the program Syllabus Plus Enterprise Timetable using the allocation option "allocate randomly" (see Figure 5.5 in the appendix). Since the academic year 2010/11 all Bachelor sections are stratified by nationality with the computer program SPASSAT.[59] Some Bachelor courses are also stratified by exchange student status. After the assignment of students to sections, the sections are assigned to time slots and the program Syllabus Plus Enterprise Timetable indicates scheduling conflicts.[60] Scheduling conflicts arise for about 5

58. About 5.6 percent of students register late. The number of late registrations in the previous year determines the number of slots that are left unfilled initially by the scheduler.
59. The stratification goes as follows: the scheduler first selects all German students (which are not ordered by any observable characteristic) and then uses the option "Allocate Students set SPREAD" which assigns an equal number of German students to all classes. Then the scheduler repeats this with the Dutch students and lastly distributes the students of all other nationalities to the remaining spots.
60. There are four reasons for scheduling conflicts: (1) the student takes another regular course at the same time. (2) The student takes a language course at the same time. (3) The student is also teaching assistant and needs to teach at the same time. (4) The student indicated non-availability for evening education. By default all students are recorded as available for evening sessions. Students can opt out

percent of the initial assignments. If the computer program indicates a scheduling conflict the scheduler manually moves students between different sections until all scheduling conflicts are resolved. After all sections have been allocated to time slots, the scheduler assigns teachers to the sections.[61] The next step in the scheduling procedure is that the section and teacher assignment is published. After this, the scheduler receives information on late registering students and allocates them to the empty spots. Throughout the scheduling process, the schedulers, who do not know the students, do not observe the previous grades of the students.

Only 20-25 students (less than one percent) officially switch section per period. This is only possible through a student advisor and is only allowed for medical reasons or due to conflict with sports practice for students who are on a list of top athletes.[62] Students sometimes switch their section unofficially when they have extra appointments. However, these are usually limited to one session and students rarely switch sections permanently.[63]

There are a few exceptions to this general procedure. We excluded a number of special courses and sections where the section assignment was not random. When the number of late registering student exceeds the number of empty spots, the scheduler creates a new section which mainly consist of late registering students. We excluded eight late registration sections from the analysis.[64] For some Bachelor courses there are special sections consisting mainly of repeating students. Whether a repeater section is created depends on the preference of the course coordinator and the number of repeat students. We excluded 34 repeater sections from the analysis. In some Bachelor courses students who are part of the Maastricht Research Based Learning (MARBLE) program are assigned to separate sections where they often are assigned to more experienced teacher. Students of this program are typically the highest performing students of their cohort. We excluded 15 sections that consist of

---

of this by indicating this in an online form. Evening sessions are scheduled from 6 p.m. to 8 p.m. and about three percent of all sessions in our sample are scheduled for this time slot.

61. About ten percent of teachers indicate time slots when they are not available for teaching. This happens before they are scheduled and requires the signature from the department chair.
62. We do not have a record for these students and can therefore not exclude them. However, section switching in these rare cases is mostly due to conflicts with medical and sports schedules and therefore unrelated to section peers.
63. It is difficult to obtain reliable numbers on unofficial switching. From our own experience and consultation with teaching staff we estimate that session switching happens in less than 1 percent of the sessions and permanent unofficial class switching happens in less than 1 percent of students.
64. Students who register late, for example, generally have a lower GPA and might be particularly busy/stressed in the period which the register late which also might affect their performance in this course. This might create a spurious relationship between GPA and grade which is due to factors like stress.

MARBLE students from the analysis.[65] In six courses the course coordinator or other education staff influenced the section composition.[66] We excluded these courses from our analysis. Some Master tracks have part time students. Part time students are scheduled mostly in evening classes and there are special classes with only part time students. We excluded 95 part time students from the analysis. We excluded first year first period courses of the two largest Bachelor programs (International Business and Economics) because in these courses only particular students, such as repeating student, have previous grades. Moreover, we exclude sections for which less than five students had a past GPA. For these courses the peer GPA does not reliably capture the peer quality of the students in the section. We excluded sections with more than 16 students (two percent) because the official class size limit according to scheduling guidelines is 15 and in special cases 16. Sections with more than 16 students are a result of room availability constraints or special requests from course coordinators.

After removing these exceptions, in our estimation sample neither students nor teachers, and not even course coordinators, influence the composition of the section.

5.3 Data

We obtained data for all students taking courses at the SBE during the academic years 2009/2010, 2010/2011 and 2011/2012. Scheduling data was provided by the Scheduling Department of the SBE. The scheduling data include information on section assignment, the allocated teaching staff, information on which day and time the sessions took place as well as a list of late registrations for our sample period. In total we have 7,746 students, 450 courses, 3,928 sections and 41,749 grades in our estimation sample. Panel A of Table 5.1 provides an overview of courses, sections and students in the different years.[67]

65. We identified pure late registration classes, repeater classes and MARBLE classes from the data. The scheduler confirmed the classes which we identified as repeater classes. The algorithm by which we identified late registration classes and MARBLE classes is available upon request.
66. The schedulers informed us about these courses.
67. We refer to each course-year combination as separate course. That means that we count a course with the same course code that takes place in three years as three separate courses.

Table 5.1 Descriptive Statistics

Panel A

Academic year	Number of courses	Number of unique students	Number of sections	Average number of students per section	Number of grades
2009 / 10	120	3,858	1,154	13.21	12,090
2010 / 11	162	4,069	1,449	13.07	14,721
2011 / 12	168	4,192	1,325	14.15	14,938
All years	450	7,746	3,928	13.49	41,749

Panel B

	Obs.	Mean	S.D.	Min	25p	Median	75p	Max
Student level information								
Course dropout	45,534	0.08	0.28	0.00	0.00	0.00	0.00	1.00
Grade first attempt	41,749	6.57	1.88	1.00	6.00	7.00	8.00	10
Final grade	45,534	6.79	1.20	1.00	6.13	6.99	7.58	10
GPA	45,501	6.57	1.37	1.00	5.70	6.75	7.52	10
Section level information								
Number of students registered for section	45,534	13.49	1.33	5.00	13.00	14.00	14.00	16.00
Number of students that dropped class	45,534	1.23	1.42	0.00	0.00	1.00	2.00	10.00
Peer GPA (based on final grades)	45,534	6.76	0.47	4.90	6.44	6.78	7.10	8.75
Peer GPA (based on first sit grades)	45,501	6.53	0.51	3.78	6.20	6.54	6.88	8.75
Within section standard deviation of peer GPA	45,506	0.91	0.32	0.00	0.67	0.89	1.12	3.10
Student Background information								
Age	42,147	20.79	2.21	16.19	19.2	20.48	22.04	41.25
Female	42,147	0.37	0.48	0.00	0.00	0.00	1.00	1.00
Dutch	45,534	0.31	0.46	0.00	0.00	0.00	1.00	1.00
German	45,534	0.49	0.50	0.00	0.00	0.00	1.00	1.00
Bachelor student	45,534	0.79	0.41	0.00	1.00	1.00	1.00	1.00
BA International Business	45,534	0.39	0.49	0.00	0.00	0.00	1.00	1.00
BA Economics	45,534	0.28	0.45	0.00	0.00	0.00	1.00	1.00
Exchange student	45,534	0.07	0.25	0.00	0.00	0.00	0.00	1.00

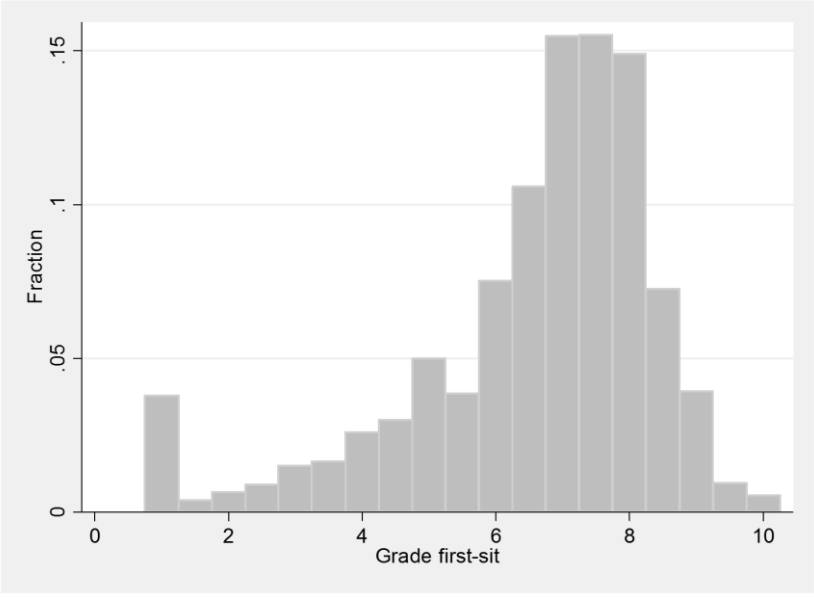
Note: This table shows the descriptive statistics of the estimation sample.

The data on student grades and student background, such as gender, age and nationality were provided by the Examinations Office of the SBE. The Dutch grading scale ranges from 1 to 10, with 5.5 being the lowest passing grade. Figure 5.1 shows the distribution of final grades in our estimation sample. The final course grade is often calculated as the weighted average of multiple graded components such as the final exam grade, participation grade, presentation grade or midterm

paper grade. The graded components and their respective weights differ by course, with most courses giving most of the weight to the final exam grade. For some courses part of the final grade consists of group graded components such as a group paper or a group presentation, for which all members of the group receive the same grade.

The influence of these group grades on the final course grade might be one of the channels through which peers affect grades. Unfortunately, we only observe the final grade and not its individual components. If the final course grade of a student after taking the final exam is lower than 5.5, the student fails the course and has the possibility to take a second attempt at the exam. We observe final grades after the first and second attempt separately.

Figure 5.1 Distribution of Grades After the First Examination



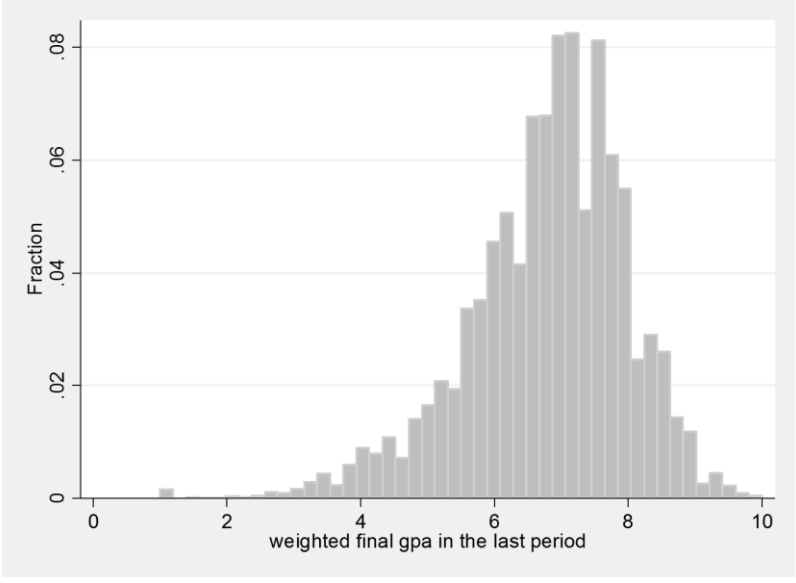
For our analysis we only use the final grade after the first exam attempt as an outcome measure, since first and second attempt grades are not comparable.[68] For the construction of the student GPA we use the final grades after the last attempt.[69]

68. The second attempt exam usually takes place two months after the first exam.
69. We decided to use the GPA calculated from final grades because this is closer to the popular understanding of GPA.

Panel B of Table 5.1 shows some descriptive statistics for our estimation sample. Our sample contains 45,534 student course registrations. Out of these 3785 (8 percent) dropped out of the course throughout the course period. We therefore observe 41,749 course grades after the first sit. Dropping out of a course means that a student was registered for a course but did not receive a grade. The average course grade after the first attempt is 6.57. About one fifth of the graded students obtain a course grade lower than 5.5 after the first attempt and therefore fail the course. The average final course grade (including grades from second and third time attempts) is 6.79, and the average GPA is 6.57. Figure 5.2 shows the distribution of the GPA based on final grades.

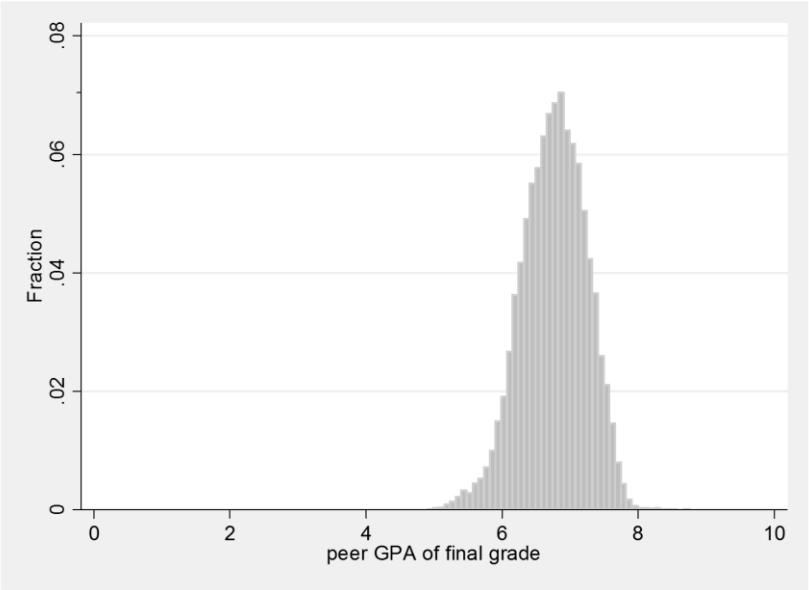
The peer GPA is the section average GPA excluding the grades of the student of interest.[70] The mean of the section level standard deviation of the GPA is 0.91. For 93 percent of our sample we know the age, gender and nationality of the students.[71] Figure 5.3 shows the distribution of peer quality measured as the average past GPA of all other students in the section.

Figure 5.2 Distribution of Own GPA



70. For a more detailed explanation, see Section 5.5 where we describe our empirical strategy.
71. We do not know age, gender and nationality for most of the exchange students.

Figure 5.3 Distribution Peer GPA



5.4 Test for Random Assignment of Students to Sections

The scheduling procedure we describe in Section 5.2.2 shows that section assignment is random. Nevertheless, we want to provide a test on whether section assignment has the properties which one would expect under random assignment. In the spirit of standard randomization checks in experiments, we test whether section dummies jointly predict student pre-treatment characteristics when controlling for scheduling and balancing indicators. The pre-treatment characteristics we look at are GPA, age, gender, and student ID rank.[72] More specifically, for each course in our sample we perform an F-test of joint significance of the section dummies in a regression of student pre-treatment characteristic on section dummies and scheduling and balancing controls. Under conditional random assignment the p-values of these F-tests should be uniformly distributed with mean 0.5 (Murdoch, Tsai, & Adcock, 2008). Furthermore, if students are randomly assigned to sections within each course, the F-test should reject the null hypothesis of no relation

72. At Maastricht University, ID numbers are increasing in tenure at the university. ID rank is the rank of the ID number. We use ID rank instead of actual ID because the SBE recently added a new digit to the ID numbers, which creates a discrete jump in the series.

between section assignment and students' pre-treatment characteristics at the 5 percent, 1 percent and 0.1 percent significance level in close to 5 percent, 1 percent and 0.1 percent of the cases, respectively.

The results of these randomization tests confirm that the section has the properties one would expect under random assignment (we provide a more detailed description on our randomization check in the appendix). The average of the p-values of the F-tests does not significantly differ from 0.5 (see Table 5.5 in the appendix) and the p-values are roughly uniformly distributed (see Figure 5.6 in the appendix). Table 5.2 shows in how many cases the F-test actually rejected the null hypothesis at the respective levels. Column (1) shows the total number of courses for each pre-treatment characteristic. Column (3) shows that the actual rejection rates at the 5 percent level are close to the expected rejection rates under random assignment. The F-tests for the regressions with the dependent variables GPA, age and ID rank are rejected slightly more often than 5 percent, the rejection rates for the dependent variable gender is slightly less than 5 percent. Columns (5) and (7) show the actual rejection rates at the 1 percent and 0.1 percent level. Also these rejection rates as a whole are close to the expected under random assignment, with the exception of age where the rejection rates is only slightly higher than we expected. All together, we present strong evidence that section assignment in our estimation sample is random, conditional on scheduling and balancing indicators.

Table 5.2 Randomization Check of Section Assignment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable	Total number of courses	Number significant	Percent significant	Number significant	Percent significant	Number significant	Percent significant
Joined F-test significant:		at the 5 percent level		at the 1 percent level		at the 0.1 percent level	
GPA	433	26	6.00%	8	1.85%	1	0.23%
Age	428	26	6.07%	11	2.57%	4	0.93%
Gender	424	17	4.01%	3	0.71%	0	0.00%
ID rank	433	22	5.08%	8	1.85%	2	0.46%

Note: This table is based on separate OLS regressions with past GPA, age, gender and ID rank as dependent variables. Explanatory variables are a set of section dummies, dummies for other course taken at the same time, and dummies for day and time of the sessions, German, Dutch, exchange student status and late registration status. Column (1) shows the total number of separate regressions. Columns (2), (4) and (6) show in how many regressions the F-test rejected the null and the 5 percent, 1 percent and 0.1 percent level respectively. Columns (3), (5) and (7) show what percentage of the regressions the F-test rejected the null at the respective levels. Differences in number of courses are due to missing observations for some of the variables.

5.5 Empirical Strategy

When trying to estimate peer effects, there are two main empirical challenges. The first challenge is the so-called reflection problem, as defined by Manski (1993), which states that one cannot disentangle the effect from the peers on the student from the effect of the student on the peers if student and peer outcomes are both determined simultaneously. In order to avoid the reflection problem we follow the standard approach in the recent peer literature and use pre-treatment measures for student and peer ability (e.g. Carrell et al., 2009; Lyle, 2007). The second challenge is selection into peer groups. The estimation of peer effects is biased when peer group assignment is related to unobservable characteristics correlated with student outcomes. However, we do not face this selection problem because the assignment of sections is random conditionally on scheduling constraints, as we have shown in Section 5.4.

We use the following model to estimate the effect of peers on grades:

$$Y_{ist} = \alpha + \beta_1 \overline{GPA}_{s-i,t-1} + \beta_2 GPA_{i,t-1} + \gamma' Z_{ist} + \varepsilon_{ist} \quad (2)$$

The dependent variable Y_{ist} is the grade of student i, in a course-specific section s, at time t. α is a constant; $\overline{GPA}_{s-i,t-1}$ is the average past GPA of all the students in the section excluding student i, $GPA_{i,t-1}$ is the past GPA of student i; Z_{ist} is a matrix of additional controls and ε_{ist} is an error term with the usual properties. Note that both $\overline{GPA}_{s-i,t-1}$ and $GPA_{i\,t-1}$ are constructed using only grades before period t in order to avoid the reflection problem. We standardized $GPA_{i\,t-1}$ and $\overline{GPA}_{s-i,t-1}$ over the estimation sample to simplify the interpretation of the coefficients.[73] In all specifications, Z_{ist} consist of dummies for day and time of the sessions, German, Dutch, exchange student status and late registration status, and year-course-period fixed effects.[74] The year-course-period fixed effects control for mean differences in outcomes across courses and time. This takes into account different grade levels in different years and courses with differing degrees of difficulty. In other specifications we also include other-course fixed effects – i.e. fixed effects for the

73. The standardized coefficients of own past GPA should be interpreted with caution because
74. For some sections the time and day of the sessions were missing. We include separate dummies for these missing values.

other course taken at the same time - and teacher fixed effects.[75] Conceptually, including scheduling controls and other-course fixed effects should pick up all non-random variation in section assignment that is due to conflicting schedules. Including stratification controls should increase the precision but not affect the size of the estimates. Teacher fixed effects should control for potential non-random assignment of teachers to sections. To allow for correlations in the outcomes of students within each course, we cluster the standard errors at the course level.

5.6 Results

Table 5.3 shows the results of OLS regressions with the standardized grade as the dependent variable.

Table 5.3 Student Grades and Peer Quality (OLS)

	(1) Std. Grade	(2) Std. Grade	(3) Std. Grade	(4) Std. Grade
Standardized peer GPA	0.0099*	0.0103*	0.0114**	0.0116**
	(0.005)	(0.006)	(0.005)	(0.006)
Standardized GPA	0.5507***	0.5505***	0.5522***	0.5520***
	(0.016)	(0.016)	(0.016)	(0.016)
Observations	41,749	41,749	41,749	41,749
R-squared	0.432	0.440	0.448	0.455
Course FE	YES	YES	YES	YES
Teacher FE	NO	YES	NO	YES
Other course FE	NO	NO	YES	YES

Note: Robust standard errors clustered at the course level are in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the standardized course grade. All specifications include dummies for day and time of the sessions, German, Dutch, exchange student status and late registration status.

The table shows that being assigned to section peers with a higher GPA causes higher course grades. The coefficient of standardized peer GPA is small but statistically significant in all models. The inclusion of teacher fixed effects and

75. Other-course fixed effects are dummies for the other course taken in the same period. These are only defined for students who take up to two courses per period. In only 1.5% of the cases students were scheduled for more than two courses and these students drop out of our sample when we include other-course fixed effects. Teacher fixed effects are fixed effects of the first teacher assigned to a session. For less than 0.4 percent students have a second teacher assigned to one section. For these teachers we use the first teacher.

other-course fixed effects hardly changes the effect size or its standard errors. The estimates reported in our preferred and most conservative specification in column (4) shows that being assigned to peers with a one standard deviation higher GPA increases the student's grade by 1.16 percent standard deviations. This effect size is about 2 percent of the effect of own GPA. In terms of the Dutch grading scale this means that, for example, an increase of peer GPA from 6.5 to 7.0 is associated with a grade increase from 6.50 to 6.521, a small and economically insignificant effect.

The specification in Table 5.3 is linear-in-mean, which implicitly assumes that all students are linearly affected by the mean ability of their peers. However, previous studies have shown that peer effects are likely heterogeneous by both student and peer ability (e.g. Burke & Sass, 2013; Carrell et al., 2013). We test for these two sources of heterogeneity simultaneously by estimating a two way interaction model similar to Carrell et al. (2013) and Burke and Sass (2013). To do this, we classify students as high GPA, middle GPA and low GPA if their GPA is in the top, middle or bottom third of the course GPA distribution respectively. We further calculate for each section the fraction of peers with high and low GPA. We then include six interactions of students' own type (high, middle and low) with fraction of high GPA and fraction of low GPA peers. Interactions with fraction of middle GPA peers are excluded because of collinearity. The interaction "High GPA * Fraction of high GPA peers", for example, can then be interpreted as the effect for high GPA students of increasing the fraction of high GPA peers in the section while keeping the fraction of low GPA peers constant. Or, put differently, the coefficient shows how high GPA students are affected if middle GPA peers (baseline) are replaced with high GPA peers.

Table 5.4 shows the coefficients of the six interactions. Overall, these effects are small in magnitude. For example, the coefficient "High GPA * Fraction of low GPA peers" suggests that an increase of 20 percent of low GPA peers, which is equivalent to replacing three out of 15 middle with low GPA peers, decreases the grade of a high GPA students by 1.9 percent standard deviations. The results for high and middle GPA students are in line with the linear-in-mean model. High and middle GPA students are positively affected by high GPA peers and negatively affected by low GPA peers. This can be seen from the sign of the interaction terms. The coefficients of the interactions with fraction of high GPA peers are positive and the coefficients of the interactions with the fraction of low GPA peers are negative. The effect of increasing the fraction of high GPA peers is (marginally) significantly different from the effect of increasing the fraction of low GPA peers for high GPA students (F-test, $p = 0.000$) and middle GPA students (F-test, $p = 0.052$). The sizes

of the effects of increasing the share of high and low GPA peers are not significantly different for high and middle ability students.

The results for low GPA students, however, are substantially different. The point estimates suggest that low GPA students are *negatively* affected by high GPA peers and negatively affected by low GPA peers. The point estimates of the interactions of low GPA student with high and low GPA peers are negative. The difference between these two estimates is not statistically significant (F-test, $p = 0.692$). The effect of increasing the fraction of high GPA peers is statistically different for low GPA compared to high and middle GPA students. There are no significant differences in the effect of increasing the share of low GPA peers.

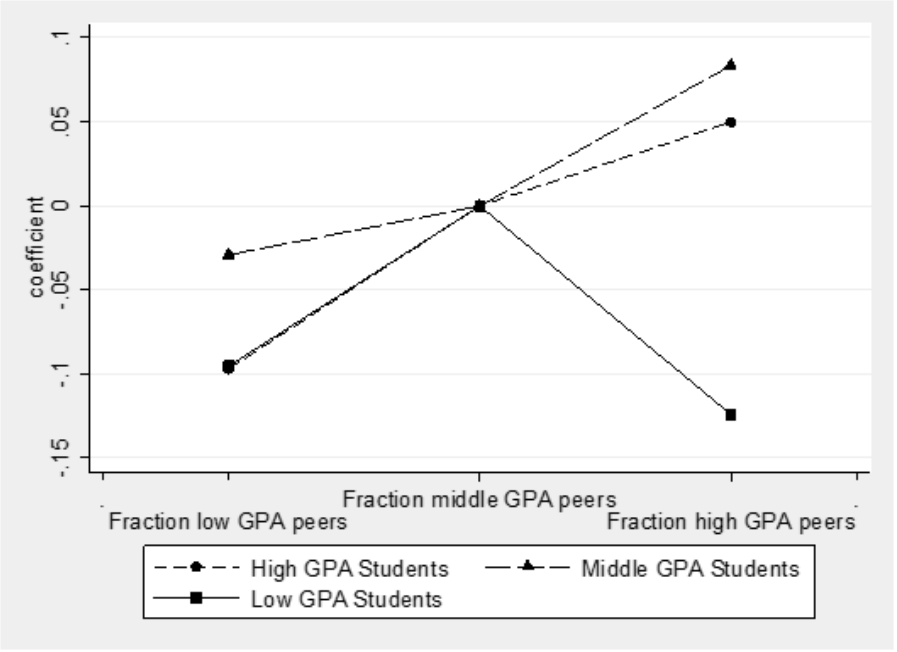
These results point to an inverse U-shape relationship between peer quality and students' own performance. Students seem to benefit from better peers as long as the difference between own and peer ability is not too large. To visualize this relationship we plot the coefficients of the interactions in Table 5.4 in Figure 5.4. The figure shows how an increase in the fraction middle and high GPA peers (compared to the fraction of middle GPA peers) affects the students' grades depending on their own ability, measured as high, middle and low GPA. Increasing the fraction of low GPA peers in the section has a negative effect on the grades for high, middle and low GPA students while increasing the fraction of high GPA peers has a positive effect for high and middle GPA students and a negative effect for low GPA students.

Table 5.4 Heterogeneous Effects

	Std. Grade
High GPA * Fraction of high GPA peers	0.0500
	(0.051)
High GPA * Fraction of low GPA peers	-0.0969**
	(0.048)
Middle GPA * Fraction of high GPA peers	0.0836
	(0.051)
Middle GPA * Fraction of low GPA peers	-0.0291
	(0.050)
Low GPA * Fraction of high GPA peers	-0.1241*
	(0.073)
Low GPA * Fraction of low GPA peers	-0.0952
	(0.067)
Observations	41,735
R-squared	0.459
F all peer variables	3.31
p-value	[0.003]
F fraction of high peers [high vs middle]	0.217
p-value	[0.641]
F fraction of high peers [high vs low]	4.015**
p-value	[0.046]
F fraction of high peers [middle vs low]	4.971**
p-value	[0.026]
F fraction of low peers [high vs middle]	1.066
p-value	[0.302]
F fraction of low peers [high vs low]	0.000
p-value	[0.984]
F fraction of low peers [middle vs low]	0.634
p-value	[0.426]

Note: Robust standard errors clustered at the course level are in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the standardized course grade. Additional controls include Std. GPA as well as dummies for high GPA, low GPA, course, teacher, other course, day and time of the sessions, German, Dutch, exchange student status and late registration status.

Figure 5.4 The Effect of Increasing Fractions of High, Middle and Low GPA Peers for Students with High, Middle and Low GPA



Note: The data points in this figure are taken from Table 5.3 using the fraction of low GPA peers as a reference category. The point estimates for "Fraction of high GPA peers" is statistically different between low GPA versus middle GPA and low GPA versus high GPA (see F-tests Table 5.3).

At the university level, our findings relate to the results of Carrell et al. (2013) who study peer effects at West Point military academy. They find that low ability students benefited from being randomly assigned to a squadron (military living community) with a high fraction of high ability peers. Based on these results, Carrell et. al (2013) then set out to maximize performance of low ability students by experimentally assigning them to squadrons with a large fraction of high ability peers while assigning middle ability students to squadrons with predominantly middle ability peers. Contrary to their own pre-treatment results, but in line with our findings, they find that this intervention negatively affected performance of low ability students. Burke and Sass (2013) studied heterogeneous peer effects at the

primary and secondary school level.[76] Their results for low ability students are in line with ours. Their estimates suggest that across school type (elementary school, middle school and high school) and subjects (mathematics and reading) low ability students are negatively affected by high and low ability peers. High ability students across all school types and subjects seem to, contrary to our results, benefit from low ability peers and, in line with our results, benefit from high ability peers. The results for middle ability students are mixed across school type and subject.

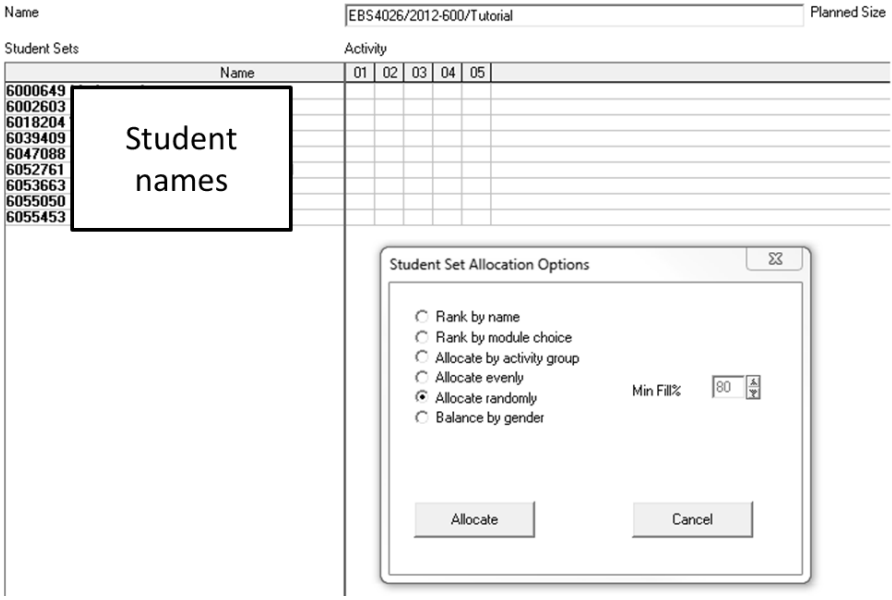
5.7 Conclusion

This article investigates peer effects in a large sample of university students where assignment to sections within a course is random conditional on scheduling constraints. Consistent with previous research we find small in size but statistically significant peer effects on student grades in the linear-in-means specification. These average effects hide some heterogeneity. While the high and middle ability students benefit from better peers, low ability students are negatively affected by both high and low ability peers compared to the middle ability baseline. These results suggest that, at least in our setting, peer effects are inverse U-shaped with respect to own ability: students seem to benefit from better peers as long as the difference between own and peer ability is not too large.

76. Note, however, that it is difficult to compare peer effects across different contexts because of differences in student ability distributions. Because it is mostly high ability students that go on to study at universities the student ability distribution is likely more narrow at the university than at the school level. Hence, the difference in ability between students which Burke and Sass (2013) classify as high and low ability at the primary and secondary level is likely larger than the differences between students which we classify as high and low ability.

5.8 Appendix

Figure 5.5 Screenshot of the scheduling program used by the SBE Scheduling Department



Note: This screenshot shows the program Plus Enterprise Timetable©.

Randomization Check

We use the following empirical specification for our tests. Take y_i as a $1xN_i$ vector of pre-treatment characteristics of students in course i. The pre-treatment characteristics we look at are GPA, age, gender, or student ID rank. $T = (t_1, \dots t_n)$ is a nxN_i matrix of section dummies (where each of the section dummies t_k is of size $1xN_i$), Z is a matrix which includes dummies for other course taken at the same time, and dummies for day and time of the sessions, German, Dutch, exchange student status and late registration status, and ε_i a $1xN_i$ vector of zero-mean independent error terms.

Our randomization tests consist of running, for each course, the following regression:

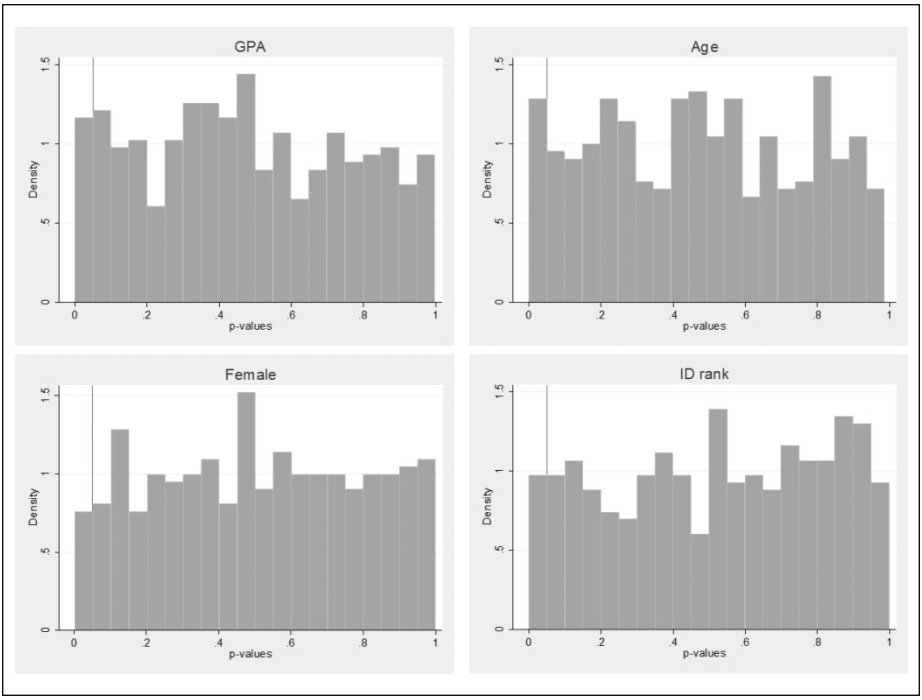
$$y_i = \alpha + T\beta + Z\gamma + \varepsilon_i \qquad \text{(A1)}$$

Under the null-hypothesis of (conditionally) random assignment to sections within each course, $\beta_1 = \beta_2 = \cdots = \beta_n = 0$. That means that the section assignment does not systematically relate to students' pre-treatment characteristic holding scheduling and stratification indicators constant. Therefore, we expect the F-test to be significant at the 5 percent level in around 5 percent of the cases, at 1 percent in around 1 percent of the cases, and at 0.1 percent in around 0.1 percent of the cases. Table 5.2 in Section 5.4 shows that the actual rejection rates are close the rejection rates expected under random assignment.

In order to investigate this issue more closely we also look at the distribution of p-values. Under the null hypothesis of conditionally random assignment, we would expect the p-values of all the regressions to closely fit a U[0,1] uniform distribution with a mean of 0.5 (Murdoch et al., 2008). Figure 5.6 shows histograms of the p-values of all four specifications, all of which are roughly uniformly distributed.

Column (2) of Table 5.5 show the mean of the p-values over all regressions reported in Table 5.2. The mean of the p-values ranges from 0.48 to 0.52. In column (3) we report p-values of a t-test, which tests whether the mean of the p-values reported in column (2) is significantly different from 0.5. None of the four t-tests rejects the null hypothesis.

Figure 5.6 Distribution of F-test P-values of β from Equation (A1) as Reported in Table 5.5



Note: These are histograms with p-values from all the regressions reported in Table 5.2.The vertical line in each histogram shows the 0.05 significance level.

Table 5.5 Randomization Check: Mean P-values

	(1)	(2)	(3)
Dependent variable:	Total number of courses	Mean of p-value	p-value of t-test (H0: Mean = 0.5)
GPA	431	0.48	0.11
Age	426	0.48	0.11
Gender	422	0.51	0.39
ID rank	432	0.52	0.12

Note: This table is based on the regressions reported in Table 5.2. Column (2) show the means of the p-values, C olumn (3) shows the p-values of t-tests testing whether the mean of the p-values in Column (3) is equal to 0.5

Chapter 6
Summary of Main Findings

6.1 Summary of Main Findings

In this thesis I have uncovered hidden relationships in the domains of overconfidence, discrimination and peer effects in education.

Chapters 2 and 3 are about the Dunning-Krueger effect (DK effect) which states that the low skilled tend to be overconfident while the high skilled are more accurate in assessing their skill. Chapter 2 is a methodological discussion on how to estimate the DK effect. We show how the fact that performance and overestimation - the respective measures of skill and overconfidence - contain measurement error can lead to biases in estimating the DK effect. This measurement error can lead to an inverse relationship between performance and overestimation even if there is no systematic relationship between skill and overconfidence. Although this problem has been recognized in the scientific literature we show that the currently used estimation methods, the split sample method and the reliability adjustment, still lead to biased estimates of the DK effect. We further show that the DK effect can be estimated consistently with the instrumental variable method using an independent performance measure as an instrument for skill. In Chapter 3 we estimate the DK effect consistently using the instrumental variable methods. In the context of students' exam grade predictions we find that the low skilled are indeed overconfident while the high skilled are more accurate. This effect is large: a one (grade) point increase in skill is associated with a 0.6 (grade) points increase in overconfidence. Comparing this estimate with three estimates from other currently used estimation methods, we show that OLS as well as the reliability adjusted OLS overestimates this effect, while the split sample method underestimates it.

Chapter 4 is about the difference between discrimination and favoritism. Most of the economic literature on discrimination assumes that differences in outcomes are driven by preferences against others – exophobia. However, we argue that they can also be driven by preference for people like oneself – endophilia. We identify endophilic and exophobic preferences with a field experiment at the SBE that assigned graders randomly to students' exams with and without names from which they could infer students' gender and nationality. We argue that there is evidence for endophilia if graders treat students who match their gender/nationality more favorable when their names are visible. Conversely, there is evidence for exophobia when graders treat students who do not match their nationality less favorable when their names are visible. On average we find endophilia but no exophobia by nationality, and neither endophilia or exophobia by gender. The effect of endophilia by nationality is large; students who are graded by a grader with matching

nationality receive on average 0.17 Std. higher grades when their names are visible. Endophilia by nationality seems to be strongest for graders with low teaching evaluations and a lot of teaching experience. Interestingly, endophilia by nationality is only present for students which the grader did not know from his/her class. We identify distributions of graders' preferences for favoritism and discrimination. Further, we extend Becker's model of discrimination to include discriminatory and favoritism preferences and show that the correlation between the two matters for observed wage differentials.

In Chapter 5 we study peer effects in education. We do this with a large dataset from the SBE where students have been randomly assigned to tutorial groups. We find that being assigned to tutorial groups with higher ability, as measured by past GPA, leads to very small increases in student grades in the linear-in-means specification. An increase in peer ability from 6.5 to 7.0 (one standard deviation) is associated with a grade increase from 6.50 to 6.52. This finding hides some heterogeneity: while middle and high ability students benefit from high ability peers, low ability students benefit from middle ability students but are harmed by high ability peers. These findings point to an inverse U-shaped relationship between performance and peer ability: students benefit from better performing peers as long as the difference between own and peer ability does not exceed a certain threshold.

Chapter 7
Valorization: Policy Recommendations

Valorization: Policy Recommendations

Here I will discuss the policy implications of the three empirical chapters of this thesis. Since all of them have been done with data of students and teachers at the SBE, many policy implications are particularly relevant for the SBE and other higher education institutions.

Key Policy Recommendations

Low ability students overestimate their academic ability. Warn low ability students that they have a high probability of dropping out, or apply stricter acceptance criteria for enrolment (Chapter 3).

There is discrimination in grading. Make grading anonymous by asking students not to write their names on the exams but only their student ID number (Chapter 4).

Peer effects are small in magnitude and we don't yet understand where they come from. Therefore, the mechanisms of peer effects should be studied more before attempting to optimize peer groups (Chapter 5).

7.1 Policy Implications of Chapter 3

In Chapter 3, we showed that it is especially the low skilled students who are most overconfident while the high skilled are more accurate. This pattern is there, although the students in this study are in their second study year and they have therefore already received plenty of feedback on their academic ability in terms of previous.

Overconfidence in one's academic ability is potentially very costly if students make wrong choices based on their inflated self-assessment. At many colleges there are large dropout rates. In the US, for example, six years after starting postsecondary 35 percent of students have dropped out, 15 percent were enrolled but did not complete a degree and only 50 percent attained a degree (Radford, Berkner et al. 2010). For most students who dropped out, going to college was very costly, for themselves but especially in Europe also for the tax payer. I realize that the fact that we observe a high drop-out rate does not necessarily mean that students make mistakes; they

might still be maximizing their utility given the circumstances and the information they have. Students, however, make their decisions not based on their actual but on their perceived academic ability. Knowing that especially the low skilled students vastly overestimate their academic ability suggests that many of them make a mistake by starting college or by not realizing that they need to put in more effort to graduate.

Mistakes in university enrolment might be prevented by being more careful with the admission of students who, based on their grades in high school or standardized tests like the GMAT or GRE, have a high likelihood of dropping out. Being careful could mean that these students are warned about the danger of dropping out or that they are not accepted in the program at all. I recognize that many universities are already selective in their admission. However, the high dropout rates suggest that they are not selective enough or do not succeed to motivate the study efforts of the low-performing students. Moreover, the main motivation of universities for being selective in the admission of students is to decide who can study if there are limited spaces. I suggest that one should also consider restricting access based on prior performance, even if the university does not have limited capacity. Doing so might prevent students from making inefficient study decisions that are costly to themselves as well as to taxpayers. Filling up every available seat at a university classroom regardless of the students' ability therefore consists of a hidden transfer from taxpayers' money to the university's coffers, and is very costly in student's welfare.

7.2 Policy Implications of Chapter 4

In Chapter 4, we show that graders give students who have the same nationality as themselves higher grades when students' names on the exams are visible than when their names are not visible. We further show that this discriminatory bias is only present for graders who did not know the students from their tutorial group.

The most straightforward implication of this result is that objectivity in grading can be increased by making exam grading anonymous. Universities can do this by simply requesting students not to write their name but only their student ID number on the exam, a step which has already been adopted at a number of universities. Until such an anonymous grading policy is implemented, individual graders who are aware of this bias and wish to correct it can cover students' names, for example by putting a sticker on top of them. There are many situations in which anonymous

grading is not possible. In these situations it will be helpful to be aware that not only discrimination but also favoritism has often negative consequences for the not favored group. In many aspects grading can be considered a zero-sum game. At the SBE, for example, grades are used as criteria to assign students to spots at popular exchange universities. For this assignment, students who are not favored are automatically disadvantaged.

Furthermore, the fact that discrimination in grading is only present when the graders do not know the students has implications for reducing discrimination in job application procedures. In job applications the cover letter and C.V. of the applicant contains the names which can often be used to infer the gender and nationality of the applicant. And indeed, by randomly assigning foreign sounding names to C.V., researchers have shown that the name of the applicant influences the probability of getting invited to an interview (see for example: Bertrand & Mullainathan, 2004). This raises the question of whether discrimination could be reduced by masking the name on the applicants' C.V. and cover letters (see for experimental evidence on this question: Behaghel, Crépon, & Le Barbanchon, 2011). Whether this will be successful in reducing discrimination is, however, not obvious because the minority status of the applicant can still be inferred at the job interview at which stage the interviewer can still discriminate (cf. Goldin & Rouse, 2000). By showing that knowing the students reduces discrimination of graders, we provide evidence that getting to know the applicant in an interview might also reduce discriminatory attitudes. This suggests that making the first stage of the application anonymous can be successful in reducing discriminatory outcomes.

7.3 Policy Implications of Chapter 5

In Chapter 5 we show that, on average, students who are randomly assigned to peers with high GPA get higher grades. Having better peers, however, is not always helpful: Low ability students are harmed by high ability peers. These heterogeneous effects suggest that one can increase overall student performance with a reassignment of students to peer groups. However, all effects are small in magnitude which suggests that the gains from peer reassignment are limited.

Economists like to evaluate decisions in terms of their costs and benefits, whereas our study has only shown the benefits of peer effects. To evaluate whether a reassignment of students to peer groups is worthwhile we also need to think about the costs of such an intervention. The organizational costs of such an intervention

are in many situations probably very cheap – in our case they would be almost zero. The mechanisms which drive the estimated peer effects, might however be costly. Peer effects might, for example, be explained by higher effort provision of the student and/or the teacher. Furthermore, recent research has suggested that social dynamics and thus the peer effect might change when students are assigned to supposedly optimally designed peer groups (Carrell, Sacerdote et al. 2013). Since we only know a small part of the information necessary to make a decision – the benefits in terms of grades – it is not justified to give a policy recommendation yet. In light of the small magnitude of our estimated peer effects and how little we know about the mechanisms driving them, I suggest that the mechanism should be studied further before we start reorganizing peer groups.

Bibliography

Abrevaya, J. and D.S. Hamermesh (2012). Charity and Favoritism in the Field: Are Female Economists Nicer (to Each Other)? *Review of Economics and Statistics*, 94(1), 202-207.

Ahmed, A.M. (2007). Group Identity, Social Distance and Intergroup Bias. *Journal of Economic Psychology*, 28 (3), 324-337.

Allport, G.W. (1954). The Nature of Prejudice. Cambridge, MA, Addison-Wesley.

Altonji, J.G. and R.M. Blank (1999). Race and Gender in the Labor Market. In O. Ashenfelter and D. Card (eds.), Handbook of Labor Economics Vol. 3C, Amsterdam, Elsevier B.V., 3143-3259.

Angrist, J. and A.B. Krueger (2001). Instrumental Variables and The Search for Identification: From Supply and Demand to Natural Experiments. The Journal of Economic Perspectives, 15 (4), 69-85.

Angrist, J.D. and K. Lang (2004). Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program. The American Economic Review, 94 (5), 1613-1634.

Arulampalam, W., et al. (2007). Is There a Glass Ceiling over Europe? Exploring the Gender Pay Gap across the Wage Distribution. Industrial and Labor Relations Review, 60, 163-186.

Bagues, M.F. and B. Esteve-Volart (2010). Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment. The Review of Economic Studies, 77 (4), 1301-1328.

Becker, G.S. (1957). The Economics of Discrimination. Chicago, University of Chicago Press Economics.

Behaghel, L., Crépon, B., & Le Barbanchon, T. (2012). Do Anonymous Resumes Make the Battlefield more Even? Evidence from a Randomized Field Experiment. Unpublished Manuscript.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. The American Economic Review, 94 (4), 991-1013.

Bollen, K.A. (1989). Structural Equations With Latent Variables, Wiley.

Brunello, G., et al. (2010). Peer Effects in Higher Education: Does the Field of Study Matter? Economic Inquiry, 48 (3), 621-634.

Burgess, S. and E. Greaves (2013). Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities. Journal of Labor Economics, 31 (3), 535-576.

Burke, M.A. and T.R. Sass (2013). Classroom Peer Effects and Student Achievement. Journal of Labor Economics, 31 (1), 51-82.

Burson, K.A., et al. (2006). Skilled or Unskilled, But Still Unaware of It: How Perceptions of Difficulty Drive Miscalibration in Relative Comparisons. Journal of Personality and Social Psychology, 90 (1), 60-77.

Cain, G.G. (1986). The Economic Analysis of Labor Market Discrimination: A survey. In O. Ashenfelter and R. Layard (eds.), Handbook of Labor Economics Vol. 2, Amsterdam, Elsevier Science B.V., 693-785.

Camerer, C. and D. Lovallo (1999). Overconfidence and Excess Entry: An Experimental Approach. The American Economic Review, 89 (1), 306-318.

Cardoso, A.R. and R. Winter-Ebmer (2010). Female-Led Firms and Gender Wage Policies. Industrial and Labor Relations Review, 64 (1), 143-163.

Carrell, S.E., et al. (2009). Does Your Cohort Matter? Measuring Peer Effects in College Achievement. Journal of Labor Economics, 27 (3), 439-464.

Carrell, S.E., et al. (2013). From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation. Econometrica, 81 (3), 855-882.

Charles, K.K. and J. Guryan (2008). Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination. Journal of Political Economy, 116 (5), 773-809.

Dane, E. and M.G. Pratt (2007). Exploring Intuition and Its Role in Managerial Decision Making. Academy of Management Review, 32 (1), 33-54.

De Paola, M. and V. Scoppa (2008). Peer Group Effects on the Academic Performance of Italian Students. Applied Economics, 42 (17), 2203-2215.

Dee, T.S. (2005). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? American Economic Review, 95 (2), 158-165.

Dillingham, A.E., et al. (1994). Gender Discrimination by Gender: Voting in a Professional Society. Industrial and Labor Relations Review, 47 (4), 622-633.

Donald, S.G. and D.S. Hamermesh (2006). What Is Discrimination? Gender in the American Economic Association, 1935-2004. The American Economic Review, 96 (4), 1283-1292.

Duflo, E., et al. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. The American Economic Review, 101 (5), 1739-1774.

Dunning, D. (2011). The Dunning-Kruger Effect: On being Ignorant of One's Own Ignorance. Advances in Experimental Social Psychology, 44, 247-296.

Ehrlinger, J., et al. (2008). Why the Unskilled are Unaware: Further Explorations of (Absent) Self-insight Among the Incompetent. Organizational Behavior and Human Decision Processes, 105 (1), 98-121.

Elder, T.E., et al. (2010). Unexplained Gaps and Oaxaca–Blinder Decompositions." Labour Economics, 17 (1), 284-290.

Ferraro, P.J. (2010). Know thyself: Competence and Self-awareness. Atlantic Economic Journal, 38 (2), 183-196.

Fershtman, C., et al. (2005). Discrimination and Nepotism: The Efficiency of the Anonymity Rule. The Journal of Legal Studies, 34 (2), 371-396.

Fong, C.M. and E. F. Luttmer (2009). What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty. American Economic Journal: Applied Economics, 1 (2), 64-87.

Fryer, R. (2011). Racial Inequality in the 21st Century: The Declining Significance of Discrimination. In O. Ashenfelter and D. Card (eds.), Handbook of Labor Economics, Vol. 4B, Amsterdam, Elsevier B.V., 855-971.

Giuliano, L., et al. (2011). Racial Bias in the Manager-Employee Relationship An Analysis of Quits, Dismissals, and Promotions at a Large Retail Firm. Journal of Human Resources, 46 (1) 26-52.

Goldberg, M.S. (1982). Discrimination, Nepotism, and Long-Run Wage Differentials. The Quarterly Journal of Economics, 97 (2), 307-319.

Goldin, C. and C. Rouse (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. The American Economic Review, 90 (4), 715-741.

Greenland, S. (2000). An Introduction to Instrumental Variables for Epidemiologists. International Journal of Epidemiology, 29 (4), 722-729.

Hanna, R.N. and L.L. Linden (2012). Discrimination in Grading. American Economic Journal: Economic Policy, 4 (4), 146-168.

Haun, D.E., et al. (2000). Assessing the Competence of Specimen-processing Personnel. Lab Medicine, 31 (11), 633-637.

Hausman, J. (2001). Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left. Journal of Economic Perspectives, 15 (4), 57-67.

Hinnerich, B.T., et al. (2011). Are Boys Discriminated in Swedish High Schools? Economics of Education Review, 30 (4), 682-690.

Hoxby, C. (2000). Peer Effects in the Classroom: Learning from Gender and Race variation. NBER Working Paper, National Bureau of Economic Research.

Klayman, J., et al. (1999). Overconfidence: It Depends on How, What, and Whom You Ask. Organizational Behavior and Human Decision Processes, 79 (3), 216-247.

Krueger, J. and R.A. Mueller (2002). Unskilled, Unaware, or Both? The Better-than-average Heuristic and Statistical Regression Predict Errors in Estimates of Own Performance. Journal of Personality and Social Psychology, 82 (2), 180-188.

Kruger, J. and D. Dunning (1999). Unskilled and Unaware of it: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-assessments. Journal of Personality and Social Psychology, 77 (6), 1121-1134.

Kruger, J. and D. Dunning (2002). Unskilled and Unaware--But Why? A Reply to Krueger and Mueller (2002). Journal of Personality and Social Psychology, 82 (2), 189-192.

Lavy, V. (2008). Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment. Journal of Public Economics, 92 (10), 2083-2105.

Levitt, S.D. and J.A. List (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? The Journal of Economic Perspectives, 21 (2), 153-174.

Lyle, D.S. (2007). Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point. The Review of Economics and Statistics, 89 (2), 289-299.

Malmendier, U. and G. Tate (2005). CEO Pverconfidence and Corporate Investment. The Journal of Finance, 60 (6), 2661-2700.

Manski, C.F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. The Review of Economic Studies. 60 (3), 531-542.

Martins, P. and I. Walker (2006). Student Achievement and University Classes: Effects of Attendance, Size, Peers, and Teachers. IZA Discussion Papers, 2490.

Murdoch, D.J., et al. (2008). P-values are Random Variables. The American Statistician, 62 (3), 242-245.

Murray, M.P. (2006). Avoiding Invalid Instruments and Coping with Weak Instruments. The Journal of Economic Perspectives, 20 (4), 111-132.

Neumark, D. (1988). Employers' Discriminatory Behavior and the Estimation of Wage Discrimination. Journal of Human Resources, 23 (3), 279-295.

Parsons, C.A., et al. (2011). Strike Three: Discrimination, Incentives, and Evaluation. The American Economic Review, 101 (4), 1410-1435.

Price, J. and J. Wolfers (2010). Racial Discrimination among NBA Referees. The Quarterly Journal of Economics, 125 (4), 1859-1887.

Radford, A.W., et al. (2010). Persistence and Attainment of 2003-04 Beginning Postsecondary Students: After Six Years. Washington, U.S. Department of Education: National Center for Education Statistics.

Rivkin, S.G., et al. (2005). Teachers, Schools, and Academic Achievement. Econometrica, 73 (2), 417-458.

Ross, L., et al. (1977). The "False Consensus Effect": An Egocentric Bias in Social Perception and Attribution Processes. Journal of Experimental Social Psychology, 13 (3), 279-301.

Ryvkin, D., et al. (2012). Are the Unskilled Doomed to Remain Unaware? Journal of Economic Psychology, 33 (5), 1012-1031.

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. The Quarterly Journal of Economics, 116 (2), 681-704.

Scheinkman, J.A. and W. Xiong (2003). Overconfidence and Speculative Bubbles. Journal of Political Economy, 111 (6), 1183-1220.

Schlösser, T., et al. (2013). How Unaware are the Unskilled? Empirical Tests of the "Signal Extraction" Counterexplanation for the Dunning-Kruger Effect in Self-Evaluation of Performance. Journal of Economic Psychology, 39, 85-100.

Tor, A. (2002). The Fable of Entry: Bounded Rationality, Market Discipline, and Legal Policy. Michigan Law Review, 101 (2), 482-568.

Turner, S. (2004). Going to College and Finishing College. Explaining Different Educational Outcomes. College Choices: The Economics of Where to Go, When to Go, and How to Pay for it, University of Chicago Press: 13-62.

Wooldridge, J. M. (2002). Econometric Analysis Cross Section Panel, MIT press.

Zimmerman, D. J. (2003). Peer Effects in Academic Outcomes: Evidence from a Natural Experiment. Review of Economics and Statistics, 85 (1), 9-23.

Biography

Jan Feld (05.04.1983, Wolfsburg, Germany) is curious about the social world and appreciates logic and common sense (he is the son of an atheist engineer). He therefore found his calling in economics (Master in Behavioral Economics, Maastricht University, 2009-2010) after a detour in International Management (Bachelor, University of Flensburg, 2005-2008). Jan started his PhD in 2010 at the Research Centre for Education and the Labour Market (ROA). His research interests are as scattered as the popular science books (Freakonomics, Predictably Irrational, etc.) that got him interest in economics. Jan aims at doing research which changes the way people think and is well identified. Aside from his academic projects, Jan also worked on a number of contract research reports (e.g. World Bank report for the ministry of Labor of the Kingdom of Saudi Arabia). Jan is a thoroughly applied economist: he tries to apply lessons from economics to his daily life. This means that he consciously tries to maximize his utility and he considers himself a sophisticated hedonist. Jan maximizes his utility by playing a lot of improv theatre. As of April 2014, Jan started a Post-Doc at Gothenburg University.

ROA Dissertation Series

1. Lex Borghans (1993), *Educational Choice and Labour Market Information*, Maastricht, Research Centre for Education and the Labour Market.

2. Frank Cörvers (1999), *The Impact of Human Capital on International Competitiveness and Trade Performance of Manufacturing Sectors*, Maastricht, Research Centre for Education and the Labour Market.

3. Ben Kriechel (2003), *Heterogeneity Among Displaced Workers*, Maastricht, Research Centre for Education and the Labour Market.

4. Arnaud Dupuy (2004), *Assignment and Substitution in the Labour Market*, Maastricht, Research Centre for Education and the Labour Market.

5. Wendy Smits (2005), *The Quality of Apprenticeship Training, Conflicting Interests of Firms and Apprentices*, Maastricht, Research Centre for Education and the Labour Market.

6. Judith Semeijn (2005), *Academic Competences and Labour Market Entry: Studies Among Dutch Graduates*, Maastricht, Research Centre for Education and the Labour Market.

7. Jasper van Loo (2005), *Training, Labor Market Outcomes and Self-Management*, Maastricht, Research Centre for Education and the Labour Market.

8. Christoph Meng (2005), *Discipline-Specific or Academic? Acquisition, Role and Value of Higher Education Competencies*, Maastricht, Research Centre for Education and the Labour Market.

9. Andreas Ammermüller (2007), *Institutional Effects in the Production of Education: Evidence from European Schooling Systems*, Maastricht, Research Centre for Education and the Labour Market.

10. Bart Golsteyn (2007), *The Ability to Invest in Human Capital*, Maastricht, Research Centre for Education and the Labour Market

11. Raymond Montizaan (2010), *Pension Rights, Human Capital Development and Well-being*, Maastricht, Research Centre for Education and the Labour Market.

12. Annemarie Nelen (2012), *Part-Time Employment and Human Capital Development*, Maastricht, Research Centre for Education and the Labour Market.

13. Jan Sauermann (2013), *Human Capital, Incentives, and Performance Outcomes*, Maastricht, Research Centre for Education and the Labour Market.

14. Harald Ulrich Pfeifer (2013), *Empirical Investigations of Costs and Benefits of Vocational Education and Training*, Maastricht, Research Centre for Education and the Labour Market.

15. Charlotte Büchner (2013), S*ocial Background, Educational Attainment and Labor Market Integration: An Exploration of Underlying Processes and Dynamics*, Maastricht, Research Centre for Education and the Labour Market.

16. Martin Humburg (2014), *Skills and the Employability of University Graduates*, Maastricht, Research Centre for Education and the Labour Market.

17. Jan Feld (2014), *Making the Invisible Visible, Essays on Overconfidence, Discrimination and Peer Effects*, Maastricht, Research Centre for Education and the Labour Market.