The Impact of Culture on Outcomes in Education - Theoretical and Empirical Evidence

Universität Hamburg,
Fakultät Wirtschafts- und Sozialwissenschaften

Dissertation
Zur Erlangung der Würde einer Doktorin
der Wirtschafts-und Sozialwissenschaften
„Dr. rer. pol."
(gemäß der PromO vom 24. August 2010)

vorgelegt von
Kathrin Thiemann
aus Gronau (Westf.), Deutschland

Hamburg, den 11. Januar 2017

Thesis Committee:

Chairman: Prof Dr. Dr. Lydia Mechtenberg

First Examiner: Prof Dr. Gerd Mühlheußer

Second Examiner: Prof. Thomas Siedler, (PhD)

Third Examiner: Jun.-Prof. Dr. Jan Marcus

The disputation was held on July 14, 2017.

Acknowledgements

I first and foremost thank my supervisor Prof. Dr. Gerd Mühlheußer for his guidance and support throughout the process of the completion of this dissertation. He was always available and reliable, when I needed his time and advice.

I also thank all my colleagues from the chair of Industrial Organization, especially Dr. Steffi Pohlkamp, Niklas Wallmeier, Dr. Berno Büchel, Dr. Leonie Gerhards, Pamela Mertens and Igor Legkiy for their company and good times.

I am grateful for financial help from the WiSo Graduate School for conducting experiments at the research laboratory and from the Department of Economics for attending conferences and workshops.

The research compiled to this dissertation has greatly profited from comments by Prof. Thomas Siedler, Dr. Stefanie Pohlkamp, Niklas Wallmeier, Prof. Dr. Gerd Mühlheußer and suggestions by members of the Economics PhD Seminar at the University of Hamburg and a couple of anonymous referees.

My special thanks go to my partner, Christoph Manys, and to my parents, Agnes and Carl Thiemann, for their enduring support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A growing body of theoretical and empirical work has shown that culture matters for a variety of economic outcomes, among these are the wealth of nations, trade, political participation, the regulation of work and gender roles (e.g. Alesina and Giuliano, 2015, 2011; Guiso et al., 2009; Tabellini, 2010). Culture affects economic outcomes because of its direct impact on expectations and preferences of individuals. An economic area where culture has received no attention so far is the field of education. The importance of cultural traits of students and teachers, however, is obvious for many aspects within this field. For example, in individualist cultures students are expected to work independently without copying ideas, whereas in collectivist cultures group collaboration is normal. In the former students are expected to participate in class, whereas in the latter students only speak up in small groups. In hierarchical cultures education is teacher-centered, in equality-favoring cultures teachers expect discussion and debate from students. In certain cultures students may learn best by observing and then doing. In others students may prefer verbal or written instructions (examples from Hofstede, 1986). These examples show that there are different areas where culture matters in teaching and learning, among these the social position of students and teachers, patterns of student-teacher and student-student interaction, the relevance of the curriculum and differences in cognitive patterns. The focus of this dissertation is on cultural differences in the area of student-student interaction, especially on educational peer effects.

A society's culture does not only shape the behavior of students and teachers in the classroom, but it also shapes national institutions. In the field of education these are for instance grading and evaluation systems, systems of private vs. public schools or teacher salaries. On the other hand also culture is influenced by institutions, such that both are endogenous variables, that are possibly determined by geography, technology, wars and other exogenous historical shocks (Alesina and Giuliano, 2015). The establishment of causal effects from culture on economic outcomes is thus prone to difficulties. Of economic interest is in particular how cultural differences influence student performance and the distribution of performances in school classes. The policy relevance of economic

research on culture is yet not obvious, since culture cannot be changed in the short-run. However, national institutions and incentive programs can be designed to take account of cultural characteristics. In the field of education this means that a globally optimal solution to many controversially discussed institutional topics (e.g. teacher incentives, public vs. private schools, class size) may not exist, but may vary from country to country depending on the cultural realities. In the greater part of this dissertation we investigate whether national school designs, especially grouping policies, should be changed to better incentivize students, given cultural differences in the area of educational peer effects.

A peer effect is present if a student's educational outcome is affected by other classmates. This influence might be direct in the sense that the presence of other students affects a student's performance without changing their behavior (Epple and Romano, 2011). Or it might be indirect, as assumed in this dissertation, with students being motivated or demotivated by social comparison with their peers leading to harder or slacking study behavior. We assume that this mechanism works differently in different cultures where students vary in their preferences for social comparison and competition. Understanding the nature of peer effects is important for the design of school systems, in particular for grouping policies, i.e. the sorting of students into classes based on their ability. We differentiate between *ability grouping*, describing within-school sorting into classes with varying difficulty for different disciplines, and *ability tracking* referring to the rigid sorting of students into different schools that differ in their demands. If homogeneity of abilities in a class induces positive peer effects, ability grouping or tracking is a source of gain. If heterogeneity of abilities leads to positive peer effects, comprehensive schooling (or *mixing*) is beneficial (Benabou, 1996). The consequences of grouping policies for the performance of students are heavily discussed in economic literature. While most empirical studies reach the conclusion that grouping and tracking leads to a more unequal distribution of performances, with high-ability students gaining and low-ability students losing, no consensus is reached on the overall effect of grouping policies on average performance (see e.g. Hoffer, 1992; Argys et al., 1996; Betts and Shkolnik, 2000; Pekkarinen et al., 2009; Duflo et al., 2011; Cortes and Goodman, 2014). In this dissertation we argue that the conflicting outcomes of these studies, which work with data from different countries, are due to differences in the cultural background of students that induce differences in peer effects.

Defining and measuring culture is troublesome. Most empirical papers define culture as "those customary beliefs and values that ethnic, religious, and social groups transmit fairly unchanged from generation to generation" as originally adopted by Guiso et al. (2006). Measures of cultural traits are usually derived from questions of international

surveys and aggregated on a country level, starting with a measure of trust developed by Knack and Keefer (1997). In theoretical papers the definition of culture is less broad and a distinction between values and beliefs is made. While beliefs can be updated and manipulated, values are often modeled as preferences in utility functions that persist over time (Alesina and Giuliano, 2015). The focus of this dissertation is on the culturally differing value of *competitiveness*, that we model as preferences for social comparison and competition. Competitiveness in economics usually plays a role in experimental studies where subjects are labeled as competitive if they select into a tournament-based compensation scheme rather than into a scheme with a piece rate (see e.g. Gneezy and Rustichini, 2004; Niederle and Vesterlund, 2007). In the theoretical part of this dissertation we introduce a novel interpretation of competitiveness as a feature of students' utility functions based on research on cultural differences in teaching and learning by Hofstede (1986). Utility functions are designed as reference dependent preferences as in Tversky and Kahneman's (1979) prospect theory, with the students' reference point given by the performance of other students in class. In the empirical part we take a survey question on the appraisal of competition from the World Values Survey (WVS) (Inglehart, 2014) and aggregate the data on a country level to approximate national preferences for competitiveness. Like Guiso et al. (2006) we assume that parents pass on their values to their children, such that the national measure of culture also approximates the preferences of students.

Figure 1.1 shows our measure of competitiveness for a sample of 28 countries in correlation with the age at which students in the countries are tracked into different schools according to their ability for the first time (data from OECD, 2013a, p.78). In countries like Germany, Hungary, Uruguay or Singapore students are sorted into different schools right after primary school. In Finland, Australia or Russia students of different abilities are taught together till the end of secondary school. These different school systems are probably the result of historic shocks that led for example to almost all countries in Central and Eastern Europe practicing rigid between-school tracking. Some countries have switched from a tracked system to a comprehensive school system in the second half of the last century (Great Britain, Finland, Sweden, Poland, Spain). Studies assessing the effects of these reforms again show inconsistent results (see Galindo-Rueda and Vignoles, 2007; Meghir and Palme, 2005; Manning and Pischke, 2006; Pekkarinen et al., 2009). Figure 1.1 shows that there is a small positive correlation between competitiveness and ability tracking, but it can be seen that countries at opposite ends of the competitiveness measure practice comprehensive schooling. If there is an optimal grouping policy given a certain culture of competitiveness, some countries could thus benefit from institutional reforms.

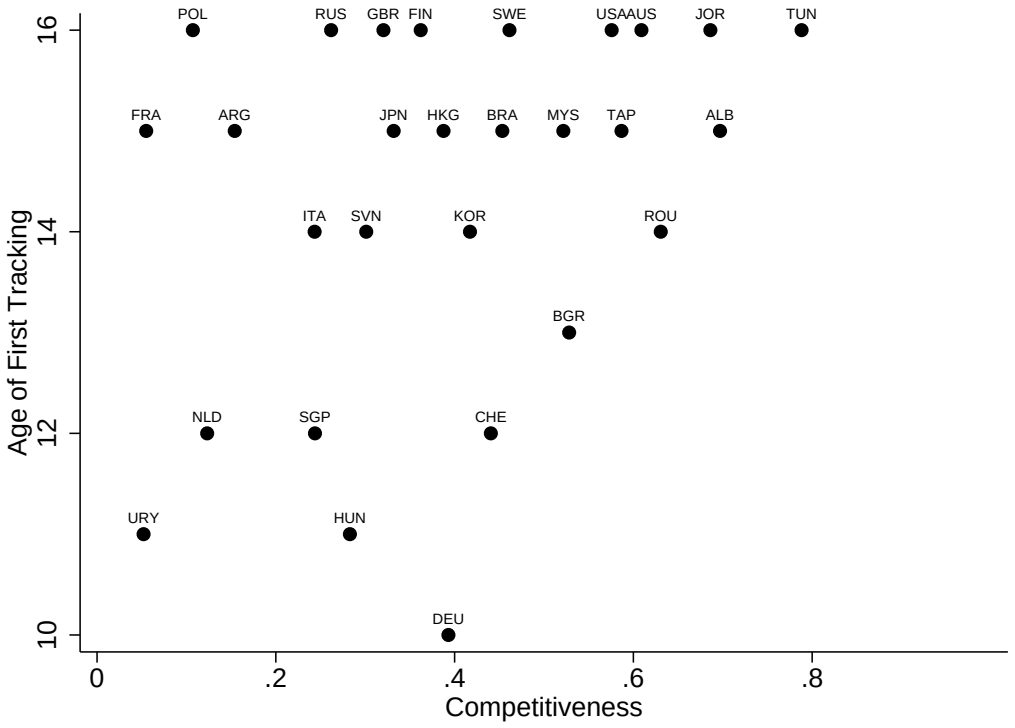


Figure 1.1: Correlation of Age of First Tracking and Competitiveness[1]

Accordingly the research question of the first two chapters of this dissertation is whether ability tracking/grouping or comprehensive schooling is to be preferred given a particular level of competitiveness. As mentioned before we use different methodologies to find an answer to this question. In Chapter 2 a theoretical model is introduced that serves as a framework for further empirical investigations. In the model students maximize their utility by choosing an optimal effort level. Students with different cultural backgrounds differ in their concern for relative position in the classroom, which is modeled by reference-dependent preferences. We contrast competitive cultures, where students compare their performance to the best performance in class, and non-competitive cultures where the reference point is the average performance. Furthermore, loss aversion with respect to the reference point is assumed to be higher in competitive cultures as well as the general weight that is put on the comparison-oriented part in the utility function. Taking into account students with heterogeneous abilities, we compare a school system where students of all abilities are taught together in one class with a system in which high-ability students are sorted into a high track and low-ability students into a low track. We show that in the Nash equilibrium in a competitive culture comprehensive schooling yields a higher average performance than ability tracking. The intuition behind

[1]The index for Competitiveness goes from 0 (least competitive country) to 1 (most competitive country). For the details of the measurement of Competitiveness refer to Section 5.3.3

this result is that low-ability students have well-performing students to look up to, which is highly motivating. In a non-competitive culture ability tracking yields a higher average performance, but also a higher dispersion of performances. This is because low-ability students lose under ability tracking compared to comprehensive schooling, but high-ability students gain more from ability tracking than low-ability students lose. In an extension we also consider the case of a participation constraint such that low-ability students stop performing if their utility becomes negative. Under this assumption ability tracking can outperform comprehensive schooling also in competitive cultures.

In Chapter 3 extensive field data from the Programme for International Student Assessment (PISA) 2012 is analyzed to test the theoretical hypotheses derived in Chapter 2. We employ an education production function approach in which student achievement is regressed on student background and school characteristic variables. As mentioned before a country-level index for competitiveness is derived from WVS data, which is interacted with an indicator for the schools policy on ability grouping from PISA. The coefficient on this interaction term yields insights on the effect of ability grouping on achievement in competitive and non-competitive cultures. Since students might self-select into schools that undertake ability grouping, an instrumental variable approach is employed, using the number of schools a school competes with as an instrument. The estimation results show that ability grouping in some or all classes increases average student achievement in competitive cultures and decreases achievement in non-competitive cultures. Employing quantile regressions we find that this holds for all students along the conditional achievement distribution, only that students at the tails are generally less affected than those closer to the median. The effect of ability grouping on the variance of achievement is not significantly different from zero in either culture. These results are different from the hypotheses developed in Chapter 2. A possible explanation for the beneficial effect of ability grouping on achievement in competitive cultures is the existence of a participation constraint and non-linear reactions to the reference point. The estimated negative effect of ability grouping on achievement in non-competitive cultures might be due to competition aversion.

To further investigate whether students behave as predicted by the model in Chapter 2, we conduct an experiment in the laboratory of which we report in Chapter 4. The caveats of the field data from PISA analyzed in Chapter 3 are that channels and preferences that drive students to perform at a certain level are hard to disentangle with the limited data from student background questionnaires. The laboratory experiment is designed to closely match the theoretical model and survey questions on loss aversion and competitiveness are included. The main component of the experiment is an effort task where subjects are

asked to solve as many multiplication problems as possible in periods of 4 minutes. Across the periods the group composition and relative performance feedback is varied. Subjects are grouped into groups of 5 either randomly or into a high or low track according to their ability. After each period subjects get feedback either on the maximum or average performance of their group. We thus do not measure a student's cultural background, but the reference point is given exogenously by the institutional design. On an aggregate level we find no significant difference in performance between random and ability grouping. However, once we distinguish by gender we find that women perform significantly better under ability grouping and male subjects under random grouping. The performance of subjects that are given the performance of the best student as a reference point is on average not higher than the performance of those that are given the average reference point, but more dispersed with more outliers at the top. However, male subjects perform significantly better when they compare themselves to the best peer instead of the average, while the opposite is true for females. In regression analysis we are able to support theory-derived hypotheses on optimal performance as being driven by loss aversion when subjects perform below their reference point. We also find evidence for peer effects evoked by the relative performance feedback, but these prove to be non-linear.

Overall we find little evidence for our hypotheses developed in Chapter 2, neither in the field nor in the laboratory. Possible explanations lie in the existence of non-linear peer effects, participation constraints, gender differences and confounding factors. Especially gender differences are important, since men and women differ in competitiveness. Hence, the experiment shows that competitive men benefit from random grouping and non-competitive women benefit from ability grouping, which corresponds to the predictions from Chapter 2.

The last Chapter 5 turns to a different crucial question in educational research, namely the search for an explanation for existing cross-country differences in intergenerational education mobility. Intergenerational education mobility refers to the relationship between parents' educational attainment and that of their children, where a high correlation describes an immobile society with little equality of opportunity. Again our focus is on cultural differences that might be potential drivers of differences between countries. We employ an econometric approach using PISA data, measuring intergenerational mobility by the effect of average years of education of the students' parents on student achievement. Measures for national culture are again derived from survey questions in the WVS about the valuation of hard work, the belief in free choice in life and, again, the appraisal of competitiveness. In the first part of this Chapter we focus on native students to compare the intergenerational education mobility among more than 40 countries. In a second part we

use data from second generation immigrants in order to overcome endogeneity problems of the cultural variables. This so-called epidemiological approach yields the advantage that the students' cultural background, which stems from their origin country as transmitted by their parents, is exogenous to the level of education mobility in the destination country. We find that disadvantages caused by family background can be overcome more easily when students come from a cultural background with high beliefs in free choice and control over their life. Competitiveness increases education mobility mainly among male students. A high cultural distance between home and host country in the valuation of hard work, however, decreases mobility among the second generation immigrants in our sample.

The central contribution of this dissertation is that we provide an explanation for cross-country differences in the effect of ability tracking on student performance and in intergenerational education mobility by considering cultural differences. To the best of our knowledge we are the first to introduce culture into a theoretical model of student effort choice and into an empirical education production function framework. We show that culture matters for outcomes in education, indicating that there cannot be international best-practices in education policies, but that policy-makers should take into account cultural aspects when they design institutions and incentive frameworks. A contribution to the literature on peer effects is that we do not only consider peer effects with respect to the average performance of a group, but also with respect to the best performance. In the laboratory experiment we show that both concepts of relative performance feedback evoke peer effects of a different type conditional on gender.

Chapter 2

Ability Tracking or Comprehensive Schooling? - A Theory on Peer Effects in Competitive and Non-Competitive Cultures[1]

Abstract

We develop a model of student decision making that shows that it depends on the culture of competitiveness in a country or region whether it is optimal to choose a school design with ability tracking or comprehensive schooling. Students with different cultural background differ in their concern for relative position in the classroom, which is modeled by reference-dependent preferences. We contrast competitive cultures, where students compare their performance with the best performance in class, and non-competitive cultures where the reference point is the average performance. Taking into account students with heterogeneous abilities, we show that the average performance in competitive cultures is maximized under comprehensive schooling and in non-competitive cultures under ability tracking. Segregation of abilities, however, always leads to a higher dispersion of performances.

JEL Codes: I28, J24, D83

Keywords: Loss Aversion, Reference Dependence, Ability Tracking, Peer Effects, Culture, Competitiveness

[1]This paper has been published as Thiemann (2017) in the Journal of Economic Behavior & Organization.

2.1 Introduction

Learning behavior of students differs to a huge extent with respect to their cultural background. In economic research that strives to determine optimal school systems and teaching practices, cultural differences in learning behavior should thus play a major role. However, culture as a determinant for outcomes in education has received little attention in economic research so far.

In this paper we are concentrating on educational peer effects, i.e. the question of how the performance of classmates influences the individual student's performance. We assume that this influence works through the channel of social comparison and competitiveness, which have been shown to vary in their extent and nature from culture to culture in various studies from psychology (e.g. Kagan and Madsen, 1971; Cox et al., 1991; Houston et al., 2005). For instance Gibbons and Buunk (1999) show in a laboratory experiment that U.S. students are significantly more comparison oriented than comparable Dutch students, measuring the time the students took to look at the performance of other participants in a computer task.

According to the cultural scientist Hofstede (1986) a competitive culture is one with high levels of social comparison, where the *best* student in class is the norm. In contrast, a non-competitive culture is one with low levels of social comparison, where students are guided by the performance of the *average* student. This behavior is on the one hand inherent to students as adopted from parents and social groups, but on the other hand influenced by teachers and institutions. Oettingen (1995, p.156) describes that teachers in competitive countries "single out high-achieving students as the ideal" and highlight their academic successes in front of the class. In line with these descriptions we set up a student-effort-choice model with reference-dependent preferences as in Kahneman and Tversky (1979). We contrast a competitive culture, where the reference point is the *best* performance in class, to a non-competitive culture, where the *average* performance is the reference point. We also assume that students are loss averse with respect to this reference point, following Hofstede who describes that for students in competitive cultures "failure in school is a severe blow to his/her self-image" and in non-competitive cultures "failure in school is a relatively minor accident" (1986, p.315). That loss aversion is significantly larger in more competitive countries, as measured by the *Masculinity* index developed by Hofstede (1984), has recently been shown by Wang et al. (2016). Conducting a survey including lottery choices in 53 countries they find, for example, a median loss aversion of 2 and 2.7 in competitive Japan and Poland respectively, as opposed to non-competitive countries like the Netherlands with 1.5 and Norway with 1.8.

The constellation of classmates, in particular whether they are of high or low ability, accordingly influences the individual student's effort choice. Therefore an important question that schools and governments face, and that shall be investigated here, is whether students should be grouped according to their ability or whether students of all abilities should be educated together. The arguments in favor of ability tracking (also referred to as streaming, phasing or ability grouping) are generally seen in the more appropriate pace of instruction. Arguments against ability tracking emphasize that it increases inequality due to the lack of positive spillovers from high achievers to low achievers. Empirical evidence on the effect of ability tracking on mean performance is mixed, while it has often been found that it indeed increases inequality (e.g. Hanushek and Woessmann, 2006; Argys et al., 1996; Hoffer, 1992). Whether the effects of ability tracking differ systematically with different cultures of competitiveness has to the best of our knowledge not been investigated so far.

In the existing literature peer effects are usually analyzed by incorporating mean ability in class in an education production function (see a literature survey by Epple and Romano, 2011). In the presence of *linear* peer effects the overall sum of students' performances is equally high when students of all abilities are taught together in one class or in classes grouped by ability. While there are no efficiency gains from ability tracking, it, however, increases inequality, since high-ability students gain from the high mean ability in the high track and low abilities suffer from the low mean ability in the low track. Differences in efficiency between ability tracking and comprehensive schooling, can be found in the presence of *non-linear* peer effects. For instance in an early paper Arnott and Rowse (1987) attempt to find a rationale for the optimal school system by maximizing a welfare function in which welfare increases in the sum of all students' final skills, but decreases with inequality. Mean ability in class here enters a Cobb-Douglas production function of students' skills, representing the peer effect. However, no clear cut recommendation on the optimal school design can be made, since results depend sensitively on the exponents in the production function.

More recent work by Benabou (1996) suggests that the peer effect (average ability) that enters the educational production function can be measured by a CES (constant elasticity of substitution) index. In the case of the elasticity of substitution tending to infinity, different abilities in the classroom are substitutes, meaning that heterogeneity of students is a source of gain. As Argys et al. (1996) have shown, comprehensive schooling then leads to efficiency gains compared with tracking. The opposite is true in the case of the elasticity of substitution approaching zero, that is when heterogeneous abilities are complements. Here heterogeneity of students is a source of loss. There are studies surveying students'

behavior suggesting that abilities rather work as complements (see Foster and Frijters, 2010), but this literature says little about what determines the elasticity of substitution. Our contribution to the existing theoretical literature is that we propose an alternative model of peer effects, which takes into account students' culture, modeled by differences in reference points and loss aversion. In contrast to the existing theoretical literature we consider peer effects as being driven not only by the average performance but also by the best performance of a group.

Among the related research is also the growing literature on loss aversion, based on original work on prospect theory by Kahneman and Tversky (1979) and Tversky and Kahneman (1991), henceforth referred to as KT (1979) and KT (1991). The phenomenon of loss aversion has been used to explain outcomes in diverse fields of research. Closest to our research is the literature referred to as "Catching up with the Joneses" (e.g. Abel, 1990; Gali, 1994). Originally used in the context of asset pricing this literature assumes that individuals get utility not only from their absolute level of income or consumption, but also from relative comparison with some social reference group. Thus, mean income or mean consumption of neighbors, peers or colleagues is incorporated into prospect theory as a reference point. Clark and Oswald (1998) develop a micro-economic model of behavior, when individuals care about relative position, i.e. they exhibit loss aversion compared with mean action in society. The model predicts herding and "following behavior", i.e. individuals follow the behavior of their reference point. In an educational setting the concept of loss aversion has been used by Levitt et al. (2016), who conduct experiments on students and find that incentives framed as losses motivate more than incentives framed as gains. To the best of our knowledge the concept of loss aversion has not been used in an educational setting with respect to the performance of classmates.

Within the framework of our effort-choice model, we compare a comprehensive school design, where students of heterogeneous abilities are together in one class with an ability tracked school design, where high abilities are sorted into a high track and low abilities into a low track. We use average performance and the variance of performances under the two systems as comparison criteria. We show that in a competitive culture a comprehensive school system provides a higher average performance. The intuition is that low ability students have well-performing students to look up to, which is highly motivating. In the case of a non-competitive culture ability tracking usually provides a higher average performance. Especially high-ability students here gain from the more homogenous peers. However, we face a trade-off between maximizing performances and minimizing inequality, since a segregation of abilities always leads to a higher dispersion of performances.

The remainder of this paper is organized as follows: Section 2.2 introduces the general

model and characterizes optimal performance. In Section 2.3 we analyze a competitive culture, where the best student's performance is the reference point, and in Section 2.4 we turn to a non-competitive culture, where the reference point is the average student's performance. Section 2.5 discusses the results and introduces some extensions. Section 2.6 concludes.

2.2 The Model

Consider a population of students who simultaneously choose their performance level p_i.[2] The utility maximization problem with reference-dependent preferences is given by:

$$\underset{p_i}{\text{Max}}\ u_i = (1-s)\cdot p_i + s\cdot v(p_i - r_i) - c(p_i, a) \tag{2.1}$$

Utility of student i depends on a linear combination of a direct private component of utility and a comparison oriented component given by the value function $v(\cdot)$. The reference point r_i is the performance student i is comparing her own performance with. The s is a parameter in the unit interval that we term the *degree of social comparison*. It is a weight on the comparison oriented utility component, whereas $(1-s)$ is a weight on the private utility component. As $s \to 0$ the standard non-behavioral model holds, where preferences are only self-interested. As $s \to 1$ only relative position matters. Costs of performing are given by $c(p_i, a) = \frac{p_i^2}{2a}$. They are increasing and convex in performance and decreasing in ability a. Students are heterogeneous in ability, thus high-ability students face lower costs than low-ability students. Specifying the value function we look at the simple linear case:

$$v(p_i - r_i) = \begin{cases} \lambda\cdot(p_i - r_i) & if \quad p_i < r_i \\ (p_i - r_i) & if \quad p_i \geq r_i \end{cases} \tag{2.2}$$

Students compare their performance with a reference point r_i. They get a positive utility as high as the difference between their own performance and the reference point if they perform higher than their reference point. Likewise they suffer from a loss in utility if they perform lower. Losses in utility are higher than the simple difference between r_i and p_i, because of loss aversion. This is captured by $\lambda > 1$, the *coefficient of loss aversion*.

The model above allows us to integrate concepts of competitiveness at three points consistent with Hofstede's (1986) interpretation of competitiveness: The first indicator is

[2]This is equivalent to an effort-choice model with the assumption that effort linearly translates into performance without noise. An example of an effort-choice model that includes a random component is Liu and Neilson (2011).

the reference point r_i. In line with Hofstede (1986) we contrast two reference points: the *average performance among the other students* (non-competitive reference point) and the *best performance among the other students* (competitive reference point). These are to be thought of as extremes on two opposite ends of possible levels of competitiveness. Second, λ the *coefficient of loss aversion* is assumed to be higher in more competitive cultures. The third indicator is s, the *degree of social comparison*. It expresses the relative importance of social comparison, which is higher in more competitive cultures and can be seen as a multiplier of the other two indicators. We assume that all students in one country have the same culture of competitiveness, i.e. they not only have the same reference point, but also have the same *loss aversion* λ and the same *degree of social comparison* s, with all indicators being higher in more competitive cultures.

2.2.1 Characterizing Optimal Performance

The marginal benefits (MB) of higher performance are different for the two cases of the value function. Assume that the reference point r_i does not depend on own performance p_i:

$$\begin{aligned} p_i < r_i: \quad MB &= (1-s) + s\lambda \equiv \mu \\ p_i \geq r_i: \quad MB &= (1-s) + s = 1 \end{aligned}$$

For ease of notation we substitute $(1-s+s\lambda) \equiv \mu \geq 1$, for the rest of the analysis, since the expression captures the joint effects of s and λ. Note that for the highest possible $s = 1$, μ is equal to the *coefficient of loss aversion* λ. For $s = 0$, i.e. in the no social comparison case, we have $\mu = 1$. Marginal costs, $MC = \frac{p_i}{a}$, are linearly increasing in performance and steeper for lower abilities. Fig. 2.1 illustrates marginal benefits and marginal costs for a given reference point r_i and different abilities $a_1 < a_2 < a_3$.

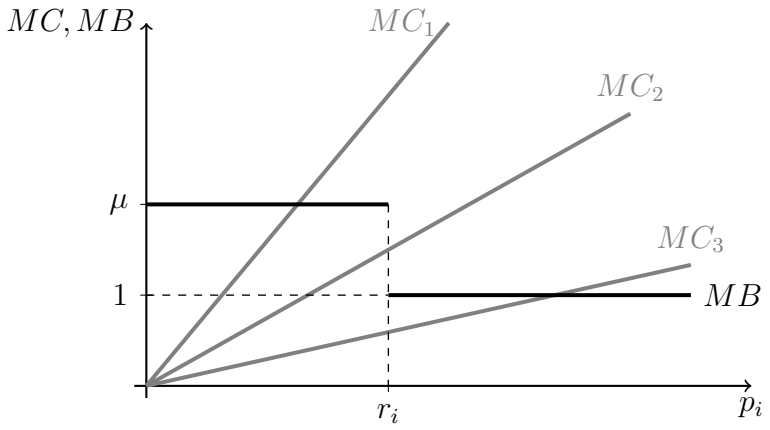


Figure 2.1: Marginal Benefits (MB) and Marginal Costs (MC) for Different Levels of Ability, with $a_1 < a_2 < a_3$

Marginal benefits equal marginal costs at $p_i = \mu a$ below the reference point and at $p_i = a$ above the reference point. The second-order condition for a local maximum is fulfilled in both cases, since $\frac{\partial^2 u}{\partial p_i^2} = -\frac{1}{a} < 0$. In the case of Fig. 2.1 students with low ability, like a_1, reach their utility maximum below the reference point and high-ability types, like a_3, optimally perform above the reference point. In the case of ability type a_2 optimal performance is found at the corner solution $p_i = r_i$. Summarizing we can characterize optimal performance as in Lemma 1.

Lemma 1.

(i) For a given ability, a, the optimal performance of student i is given by the following best response as a function of the reference point r_i:

$$BR_i(a, r_i) = \begin{cases} \mu a & if \quad a < \frac{r_i}{\mu} & \textbf{\textit{case 1}} \\ r_i & if \quad \frac{r_i}{\mu} \leq a < r_i & \textbf{\textit{case 2}} \\ a & if \quad a \geq r_i & \textbf{\textit{case 3}} \end{cases} \tag{2.3}$$

(ii) Students' best responses are non-decreasing in ability, i.e. $\frac{\partial BR_i(\cdot, r_i)}{\partial a} \geq 0$.

Since optimal performance depends on the reference point, which is given by other students optimal performances, (2.3) can be termed *best response* function. In both interior solutions the reference point does not influence the *level* of performances, but it determines in which of the three *cases* the student performs. The higher the reference point

r_i, the higher the thresholds $\frac{r_i}{\mu}$ and r_i that distinguish the cases, making a performance in case 1 for a given a more likely. In both interior solutions optimal performance is linearly increasing in ability a. Moreover, for a given reference point r_i, ability determines in which case the student performs, with a higher ability resulting in a higher case. We can thus conclude that performance is increasing in ability (see Lemma 1 (ii)). This can also be observed in Fig. 2.1. Furthermore, performance in case 1 is linearly increasing in μ, indicating that a more competitive student below the reference point performs higher. This is because of the higher marginal benefit from performance due to higher *loss aversion*.

2.3 Competitive Culture

2.3.1 Comprehensive Schooling

Assume that students are identified by their ability, which is uniformly distributed along the ability segment $a \in [\underline{a}, \bar{a}]$. The reference point of each student is given by the performance of the best student among the other students in class. Due to the continuous ability distribution we mathematically capture the reference point by the supremum of all other performances: $r_i = sup(p_{j \neq i})$. To find Nash equilibrium performances under a comprehensive school regime we analyze the case where students of all abilities $a \in [\underline{a}, \bar{a}]$ are in one class, each with the best response function as given by (2.3). Since student i is fully identified by her ability a, we index with a instead of i from now on.

Proposition 1. *In a competitive culture under comprehensive schooling a set of Nash equilibrium performances can be supported, where the student with the highest ability $\bar{a}$ performs within $p^*_{\bar{a}} \in [\bar{a}, \mu\bar{a}]$ and students with ability $a \in [\underline{a}, \bar{a})$ perform as follows:*

$$p^*_a(p^*_{\bar{a}}) = \begin{cases} \mu\ a & if\ a < \frac{p^*_{\bar{a}}}{\mu} \\ p^*_{\bar{a}} & if\ a \geq \frac{p^*_{\bar{a}}}{\mu} \end{cases} \tag{2.4}$$

Proof

(i) Since optimal performance is non-decreasing in ability (see Lemma 1 (ii)), the performance of the student with the highest ability, $p_{\bar{a}}$, must be in the set of the highest performances in class. Her performance thus always serves as a reference point for all other students.

(ii) Consider the following cases of student $\bar{a}$'s performance. The best responses of students $[\underline{a}, \bar{a})$ are given by (2.3), where $r_a = p_{\bar{a}}\ \forall\ a \in [\underline{a}, \bar{a})$.

(1) $p_{\bar{a}} < \bar{a}$:

All three cases of (2.3) may exist, since there is a $a > p_{\bar{a}}$ if $p_{\bar{a}} < \bar{a}$. The supremum of the performances of students $[\underline{a}, \bar{a})$ is then $\bar{a}$, which is the reference point of student $\bar{a}$: $r_{\bar{a}} = \bar{a}$. Student $\bar{a}$'s best response is then in the third case of (2.3), since $\bar{a} \geq \bar{a}$: $BR_{\bar{a}} = \bar{a}$. This, however, contradicts the case assumption $p_{\bar{a}} < \bar{a}$, such that there is no mutual best response in this case.

(2) $p_{\bar{a}} > \mu\bar{a}$:

In this case the best response of students $[\underline{a}, \bar{a})$ is given by case 1 of (2.3), since $a < \frac{p_{\bar{a}}}{\mu}$: $BR_a = \mu a \ \forall \ a \in [\underline{a}, \bar{a})$. The supremum of these performances is then $\mu\bar{a}$, which is the reference point of student $\bar{a}$: $r_{\bar{a}} = \mu\bar{a}$. Student $\bar{a}$'s best response is then in the second case of (2.3), since the case condition, $\frac{r_{\bar{a}}}{\mu} \leq \bar{a}$, is binding: $BR_{\bar{a}} = \mu\bar{a}$. This does again contradict the case assumption $p_{\bar{a}} > \mu\bar{a}$, such that there is no mutual best response in this case.

(3) $\bar{a} \leq p_{\bar{a}} \leq \mu\bar{a}$:

In this case the best response of students $[\underline{a}, \bar{a})$ is either in the first or second case of (2.3). The third case is not possible, since $a < p_{\bar{a}}$. The supremum of these performances, and thus student $\bar{a}$'s reference point, is then: $r_{\bar{a}} = p_{\bar{a}}$. Student $\bar{a}$'s best response is then in the second or third case of (2.3), with $BR_{\bar{a}} = p_{\bar{a}}$ if $\frac{p_{\bar{a}}}{\mu} \leq \bar{a} < p_{\bar{a}}$ and $BR_{\bar{a}} = \bar{a}$ if $p_{\bar{a}} = \bar{a}$. That means all performances that fulfill the case condition $\bar{a} \leq p_{\bar{a}} \leq \mu\bar{a}$ are possible best responses. We have thus found a class of mutual best responses, with student $\bar{a}$'s performance in the set $p^*_{\bar{a}} \in [\bar{a}, \mu\bar{a}]$ and all other students performing as in the first two cases of (2.3) with $p^*_{\bar{a}}$ as reference point.

As stated in Proposition 1 there are multiple Nash equilibria with student $\bar{a}$ setting the reference point for all others from within the interval $p^*_{\bar{a}} \in [\bar{a}, \mu\bar{a}]$. The Nash equilibrium which Pareto dominates all other equilibria is found where student $\bar{a}$ performs at the lower bound, $p_{\bar{a}} = \bar{a}$, and all other students follow this reference point. This is the equilibrium with the lowest effort costs and highest utility for all students. Equilibrium performances of students $[\underline{a}, \bar{a})$ are increasing in μ as long as the students perform below the reference point. For a high enough μ the students eventually switch to a performance at the reference point. For $\mu \geq \frac{p^*_{\bar{a}}}{\underline{a}}$ all performances have converged to a symmetric equilibrium with every student performing at $p^*_{\bar{a}}$. Fig. 2.2 illustrates the Pareto dominant equilibrium. The gray lines are equilibrium performances as functions of μ for a discrete choice of students from the ability distribution. The best student's performance is highlighted in black.

It can be seen that higher competitiveness μ drags performances up to the best perfor-

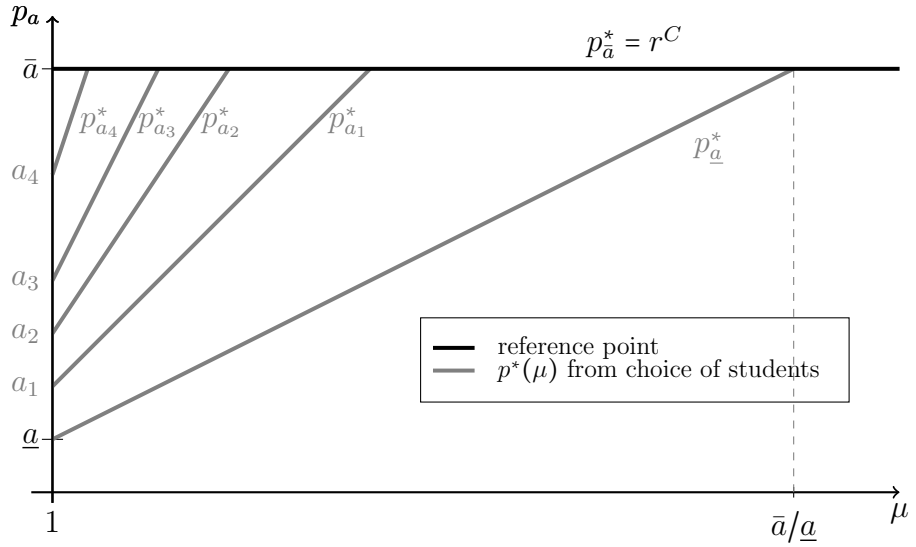


Figure 2.2: Equilibrium Performances as Functions of μ for a Discrete Choice of Students with $\underline{a} < a_1 < a_2 < a_3 < a_4 < \bar{a}$ in a Competitive Culture under Comprehensive Schooling

mance in class. This is what Clark and Oswald (1998) describe as "following behavior". The implication is that performances in more competitive, and in particular more loss averse cultures, are not as dispersed and higher than in less competitive cultures. This is because higher loss aversion can be translated into a higher motivation of students to reach the reference point. In the given case of the *best* student being the reference point, all students except the best student herself are affected by this higher return to performance.

2.3.2 Ability Tracking

Under ability tracking low-ability students $a \in \left[\underline{a}, \frac{\bar{a}+\underline{a}}{2}\right]$ are taught together in a low track and high-ability students $a \in \left(\frac{\bar{a}+\underline{a}}{2}, \bar{a}\right]$ are taught together in a high track. Assume that ability is fully observable and students are correctly assigned to tracks according to their ability.

Proposition 2. *In a competitive culture under ability tracking Nash equilibria can be described as follows:*

(i) In a high track that consists of students with ability $a \in \left(\frac{\bar{a}+\underline{a}}{2}, \bar{a}\right]$ a set of Nash equilibrium performances can be supported, where the student with the highest ability $\bar{a}$

performs within $p_{\bar{a}}^* \in [\bar{a}, \mu\bar{a}]$ *and students with ability* $a \in \left(\frac{\bar{a}+\underline{a}}{2}, \bar{a}\right)$ *perform as follows:*

$$p_a^*(p_{\bar{a}}^*) = \begin{cases} \mu\ a & if\ a < \frac{p_{\bar{a}}^*}{\mu} \\ p_{\bar{a}}^* & if\ a \geq \frac{p_{\bar{a}}^*}{\mu} \end{cases} \tag{2.5}$$

(ii) In a low track that consists of students with ability $a \in \left[\underline{a}, \frac{\bar{a}+\underline{a}}{2}\right]$ *a set of Nash equilibrium performances can be supported, where the student with the highest ability* $\frac{\bar{a}+\underline{a}}{2}$ *performs within* $p^*_{\frac{\bar{a}+\underline{a}}{2}} \in \left[\frac{\bar{a}+\underline{a}}{2}, \mu\frac{\bar{a}+\underline{a}}{2}\right]$ *and students with ability* $a \in \left[\underline{a}, \frac{\bar{a}+\underline{a}}{2}\right)$ *perform as follows:*

$$p_a^*\left(p^*_{\frac{\bar{a}+\underline{a}}{2}}\right) = \begin{cases} \mu\ a & if\ a < \frac{p^*_{\frac{\bar{a}+\underline{a}}{2}}}{\mu} \\ p^*_{\frac{\bar{a}+\underline{a}}{2}} & if\ a \geq \frac{p^*_{\frac{\bar{a}+\underline{a}}{2}}}{\mu} \end{cases} \tag{2.6}$$

Proof

(i) High track: The proof is the same as for Proposition 1, since the upper bound of the ability distribution, and hence the reference point, is the same as under comprehensive schooling. Mutual best responses are the same as in Proposition 1.

(ii) Low track: Following from Lemma 1 (ii) the highest-ability student in the low track, student $\frac{\bar{a}+\underline{a}}{2}$, is in the set of the highest performing students in the low track. Her performance thus serves as a reference point for all other students in the low track. By substituting $\bar{a}$ with $\frac{\bar{a}+\underline{a}}{2}$ in the proof for Proposition 1 the equilibrium performances as in Proposition 2 follow immediately.

We see that nothing changes for high-ability students $\left(\frac{\bar{a}+\underline{a}}{2}, \bar{a}\right]$ when switching from a comprehensive school system to an ability tracked system. Performances of low-ability students $\left[\underline{a}, \frac{\bar{a}+\underline{a}}{2}\right)$ are capped by the performance of the highest-ability student in the low track. Symmetric equilibria under ability tracking are reached for smaller levels of μ than under comprehensive schooling. A symmetric equilibrium in the high track, with all students performing at $p_{\bar{a}}^*$, is reached already when $\mu \geq \frac{2\bar{a}}{\bar{a}+\underline{a}}$ in the Pareto dominant case. In the low track a symmetric equilibrium in the Pareto dominant case with $p^*_{\frac{\bar{a}+\underline{a}}{2}} = \frac{\bar{a}+\underline{a}}{2}$ is reached if $\mu \geq \frac{\bar{a}+\underline{a}}{2\underline{a}}$. Fig. 2.3 illustrates this Pareto dominant Nash equilibrium under ability tracking for a discrete choice of students.

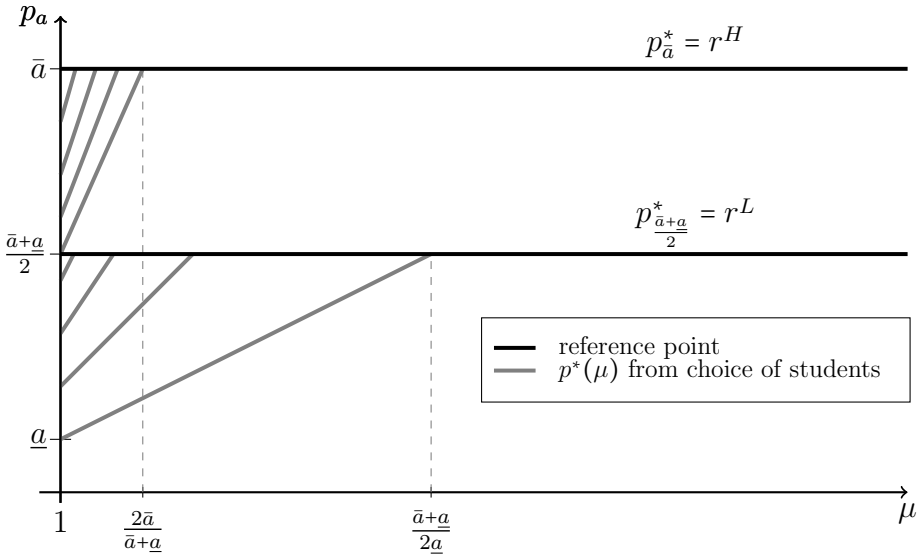


Figure 2.3: Equilibrium Performances as Functions of μ for a Discrete Choice of Students in a Competitive Culture under Ability Tracking

2.3.3 Comprehensive Schooling vs. Ability Tracking

Formally comparing the average performances, $\bar{p}^C$ and $\bar{p}^T$, and the variances of performances, σ^C and σ^T, under the two regimes in competitive cultures leads to Proposition 3. We thereby only take into account the Pareto dominant equilibria.

Proposition 3.

(i) In a competitive culture the average performance in the Pareto dominant Nash equilibrium is strictly higher under comprehensive schooling than under ability tracking for any $\mu > 1$.

(ii) The variance of performances in the Pareto dominant Nash equilibrium is strictly higher under ability tracking than under comprehensive schooling for any $\mu > 1$.

The formal proof of Proposition 3 is provided in Appendix 2.A. According to this Proposition ability tracking is never the better option in a competitive culture. The intuition behind this is that in comprehensive schools all students are affected by the motivating power of the high reference point. In an ability tracked system this high motivating force is restricted to the students in the high track. Low-ability students' performances stay at low levels, because they have no high achieving peer to look up to. Since these low-ability students do not converge to high performance levels under ability tracking, the variance of performances is higher under a segregating regime.

2.4 Non-Competitive Culture

2.4.1 Comprehensive Schooling

The non-competitive reference point is average performance among the other students in class. Since we are again considering a uniform ability distribution $a \in [\underline{a}, \bar{a}]$ with a continuum of students, the reference point of student a in a comprehensive school is given by: $r_a^C = \bar{p}^C = \frac{1}{\bar{a}-\underline{a}} \int_{\underline{a}}^{\bar{a}} p_a \, da$. We do not have to take into account that the individual student does not consider her own performance to be part of the average, since she has measure 0 in the continuum. The reference point in a comprehensive school class is thus the same for every student, given by $\bar{p}^C$.

Proposition 4. *In a non-competitive culture under comprehensive schooling Nash equilibria can be described as follows:*

(i) If $\mu \leq \frac{\bar{a}}{\underline{a}}$ the following Nash equilibrium performance p_a^* can be supported for all students with $a \in [\underline{a}, \bar{a}]$:

$$p_a^* = \begin{cases} \mu \; a & if \quad a < \frac{\bar{a}+\underline{a}\sqrt{\mu}}{\mu+\sqrt{\mu}} \\ \frac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{\sqrt{\mu}+1} & if \quad \frac{\bar{a}+\underline{a}\sqrt{\mu}}{\mu+\sqrt{\mu}} \leq a < \frac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{\sqrt{\mu}+1} \\ a & if \quad a \geq \frac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{\sqrt{\mu}+1} \end{cases} \tag{2.7}$$

(ii) If $\mu > \frac{\bar{a}}{\underline{a}}$ a set of symmetric Nash equilibria with all students performing at the same level p^* with $p^* = p_a^* \forall a \in [\underline{a}, \bar{a}]$ within the interval $p^* \in [\bar{a}, \mu\underline{a}]$ can be supported.

Proof

(i) Average performance in a comprehensive school class $\bar{p}^C$ is given by:

$$\bar{p}^C = \frac{1}{\bar{a}-\underline{a}} \left(\int_{\underline{a}}^{\frac{\bar{p}^C}{\mu}} \mu a \, da + \int_{\frac{\bar{p}^C}{\mu}}^{\bar{p}^C} \bar{p}^C \, da + \int_{\bar{p}^C}^{\bar{a}} a \, da \right) \tag{2.8}$$

For each of the three integrals to be non-negative we need to assume that $\underline{a} \leq \frac{\bar{p}^C}{\mu} \leq \bar{p}^C \leq \bar{a}$, which can be transformed to $\mu \leq \frac{\bar{a}}{\underline{a}}$. Solving (2.8) for $\bar{p}^C$ yields two solutions of which one violates this assumption,[3] such that we are left with one solution: $\bar{p}^C = \frac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{\sqrt{\mu}+1}$. This is the reference point for all students: $r_a = \bar{p}^C \forall a \in [\underline{a}, \bar{a}]$.

[3]The second solution is given by $\bar{p}_2^C = \frac{\bar{a}\sqrt{\mu}-\underline{a}\mu}{\sqrt{\mu}-1}$. It can be shown that this expression is bigger than $\bar{a}$ for all $\mu < \frac{\bar{a}}{\underline{a}}$. For $\mu = \frac{\bar{a}}{\underline{a}}$ it equals $\bar{p}^C$.

Substituting $\bar{p}^C$ into the best response function (2.3) yields equilibrium performances as in (2.7).

(ii) Substituting $\mu = \frac{\bar{a}}{\underline{a}}$ into $\bar{p}^C$ we find that average performance has reached the upper bound of the previous assumption: $\bar{p}^C = \bar{a} = r_a$. From the best response function (2.3) we see that a symmetric equilibrium is reached, where the highest-ability student performs in case 3 with performance $\bar{a}$ and all others perform in case 2, also with performance $\bar{a}$. This equilibrium exists for μ's high enough, such that also the lowest-ability student performs in case 2 while case 1 is empty, i.e. if $\underline{a} \geq \frac{r_a}{\mu} \rightarrow \underline{a} \geq \frac{\bar{a}}{\mu} \Leftrightarrow \mu \geq \frac{\bar{a}}{\underline{a}}$.

(iii) There can be more symmetric equilibria where all students perform in the second case. Average performance (the reference point) then equals individual equilibrium performance. These equilibria exist for all reference points that fulfill $a < r_a \leq \mu a$ for every student. For the highest and the lowest-ability student these intervals overlap if $\mu\underline{a} > \bar{a}$, i.e. if $\mu > \frac{\bar{a}}{\underline{a}}$. Then there are symmetric equilibria with $\bar{a} < p^* \leq \mu\underline{a}$ where $p^* = p_a^* \forall a \in [\underline{a}, \bar{a}]$.

Taking a closer look at the thresholds that distinguish the cases in equilibrium as given in Proposition 4, it can be shown that $\bar{p}^C = \frac{\bar{a}\sqrt{\mu}+a\mu}{\sqrt{\mu}+1}$ increases in μ and $\frac{\bar{p}^C}{\mu} = \frac{\bar{a}+\underline{a}\sqrt{\mu}}{\mu+\sqrt{\mu}}$ decreases in μ. Fig. 2.4 illustrates in which case the students of ability a perform, depending on μ.

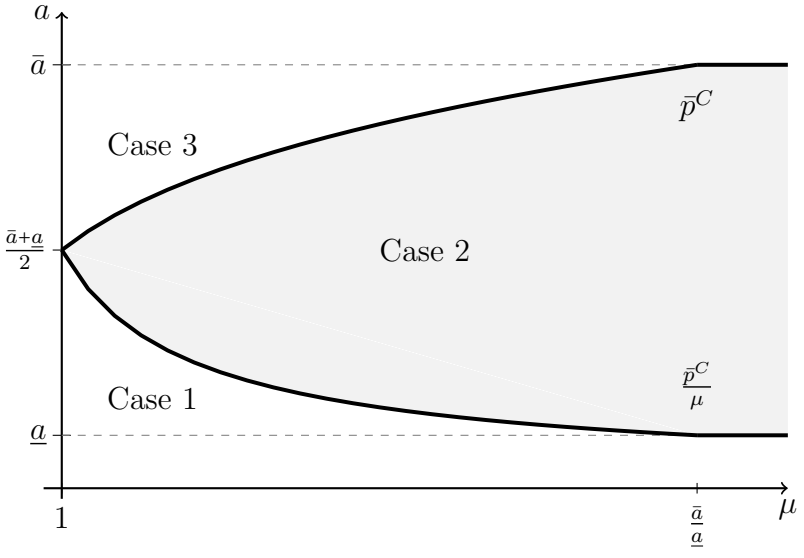


Figure 2.4: Case Thresholds for Students of Abilities $[\underline{a}, \bar{a}]$ under Comprehensive Schooling

We find the case of all students performing proportionally to their ability for $\mu = 1$. In this type of equilibrium all high-ability students $\left[\frac{\bar{a}+\underline{a}}{2}, \bar{a}\right]$ perform in the third case, i.e.

above the reference point, and all low-ability students $\left[\underline{a}, \frac{\bar{a}+\underline{a}}{2}\right)$ perform in the first case below the reference point. The bigger μ, the more students, starting with students close to the reference point, are switching to a performance at the average. This is because performance in case 1 is increasing in μ, while the threshold $\frac{\bar{p}^C}{\mu}$ decreases in μ. Average performance, which is also the threshold between case 2 and 3, increases in μ, since it encompasses performances in case 1. The fewer students perform in case 1, the flatter is the slope of average performance in μ. Because of the increasing reference point more and more students from case 3, whose performance is constant in μ, fall below this threshold and also perform at the average. For $\mu \geq \frac{\bar{a}}{\underline{a}}$ a symmetric equilibrium is reached with all students performing at the reference point. We again observe a convergence towards the highest possible performance as we have already seen in a competitive culture. Fig. 2.5 illustrates performances for some representative students depending on μ.

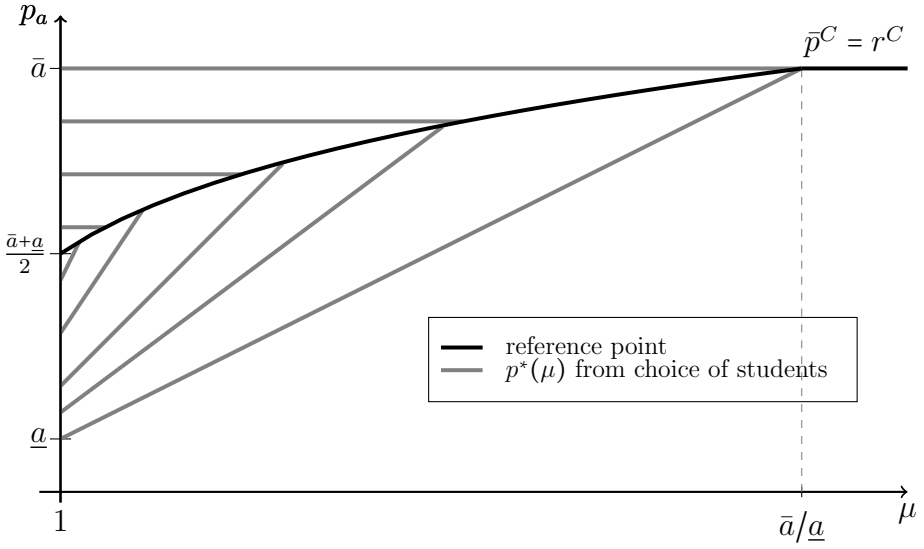


Figure 2.5: Equilibrium Performances as Functions of μ for a Discrete Choice of Students in a Non-Competitive Culture under Comprehensive Students

2.4.2 Ability Tracking

Reference points under ability tracking are the average performance in the low track $r_a^L = \bar{p}^L = \frac{2}{\bar{a}-\underline{a}} \int_{\underline{a}}^{\frac{\bar{a}+\underline{a}}{2}} p_a \, \mathrm{d}a \;\; \forall a \in \left[\underline{a}, \frac{\bar{a}+\underline{a}}{2}\right]$ and average performance in the high track $r_a^H = \bar{p}^H = \frac{2}{\bar{a}-\underline{a}} \int_{\frac{\bar{a}+\underline{a}}{2}}^{\bar{a}} p_a \, \mathrm{d}a \;\; \forall a \in \left(\frac{\bar{a}+\underline{a}}{2}, \bar{a}\right]$. Equilibrium performances as in Proposition 5 follow immediately from Proposition 4 by substituting $\bar{a}$ with $\frac{\bar{a}+\underline{a}}{2}$ for the solution of the low track and $\underline{a}$ with $\frac{\bar{a}+\underline{a}}{2}$ for the high track.

Proposition 5. *In a non-competitive culture under ability tracking Nash equilibria can be described as follows:*

(i) If $\mu \leq \frac{2\bar{a}}{\bar{a}+\underline{a}}$ *the following Nash equilibrium performance* p_a^* *can be supported for all students in the high track with* $a \in \left(\frac{\bar{a}+\underline{a}}{2}, \bar{a}\right]$*:*

$$p_a^* = \begin{cases} \mu\ a & if \quad a < \frac{2\bar{a}\sqrt{\mu}+(\bar{a}+\underline{a})\mu}{2\mu\left(\sqrt{\mu}+1\right)} \\ \frac{2\bar{a}\sqrt{\mu}+(\bar{a}+\underline{a})\mu}{2\left(\sqrt{\mu}+1\right)} & if \quad \frac{2\bar{a}\sqrt{\mu}+(\bar{a}+\underline{a})\mu}{2\mu\left(\sqrt{\mu}+1\right)} \leq a < \frac{2\bar{a}\sqrt{\mu}+(\bar{a}+\underline{a})\mu}{2\left(\sqrt{\mu}+1\right)} \\ a & if \quad a \geq \frac{2\bar{a}\sqrt{\mu}+(\bar{a}+\underline{a})\mu}{2\left(\sqrt{\mu}+1\right)} \end{cases} \tag{2.9}$$

(ii) If $\mu > \frac{2\bar{a}}{\bar{a}+\underline{a}}$ *a set of symmetric Nash equilibria with all students in the high track performing at the same level* p^* *with* $p^* = p_a^* \forall a \in \left(\frac{\bar{a}+\underline{a}}{2}, \bar{a}\right]$ *within the interval* $p^* \in \left[\bar{a}, \frac{\mu(\bar{a}+\underline{a})}{2}\right]$ *can be supported.*

(iii) If $\mu \leq \frac{\bar{a}+\underline{a}}{2\underline{a}}$ *the following Nash equilibrium performance* p_a^* *can be supported for all students in the low track with* $a \in \left[\underline{a}, \frac{\bar{a}+\underline{a}}{2}\right]$*:*

$$p_a^* = \begin{cases} \mu\ a & if \quad a < \frac{(\bar{a}+\underline{a})\sqrt{\mu}+2\underline{a}\mu}{2\mu(1+\sqrt{\mu})} \\ \frac{(\bar{a}+\underline{a})\sqrt{\mu}+2\underline{a}\mu}{2(1+\sqrt{\mu})} & if \quad \frac{(\bar{a}+\underline{a})\sqrt{\mu}+2\underline{a}\mu}{2\mu(1+\sqrt{\mu})} \leq a < \frac{(\bar{a}+\underline{a})\sqrt{\mu}+2\underline{a}\mu}{2(1+\sqrt{\mu})} \\ a & if \quad a \geq \frac{(\bar{a}+\underline{a})\sqrt{\mu}+2\underline{a}\mu}{2(1+\sqrt{\mu})} \end{cases} \tag{2.10}$$

(iv) If $\mu > \frac{\bar{a}+\underline{a}}{2\underline{a}}$ *a set of symmetric Nash equilibria with all students in the low track performing at the same level* p^* *with* $p^* = p_a^* \forall a \in \left[\underline{a}, \frac{\bar{a}+\underline{a}}{2}\right]$ *within the interval* $p^* \in \left[\frac{\bar{a}+\underline{a}}{2}, \mu\underline{a}\right]$ *can be supported.*

Behavior in the tracks is similar to that in a comprehensive school class, i.e. performances converge to the respective average performance in the tracks. Fig. 2.6 shows in which case of the best response function a student a performs depending on μ. Just like under comprehensive schooling there is an equilibrium under ability tracking where all students perform proportionally to their ability for $\mu = 1$. With μ increasing, more and more students in the high and low track perform in case 2, i.e. at the average performance of the respective track. Since average performance is equal to the upper threshold, we can observe that average performance in both tracks is converging towards the performance of the highest-ability student in the tracks. The striking difference to comprehensive schooling is that symmetric equilibria in the tracks are reached for much lower μ's. This is because the distance of own performance to the reference point is now lower for many students. Fig. 2.7 illustrates Pareto dominant equilibrium performances, where students perform at the lower bound of the symmetric equilibria. Unlike under comprehensive schooling students $\left(\frac{\bar{a}+\underline{a}}{2}, \frac{3\bar{a}+\underline{a}}{4}\right)$ find themselves below the average, which motivates them

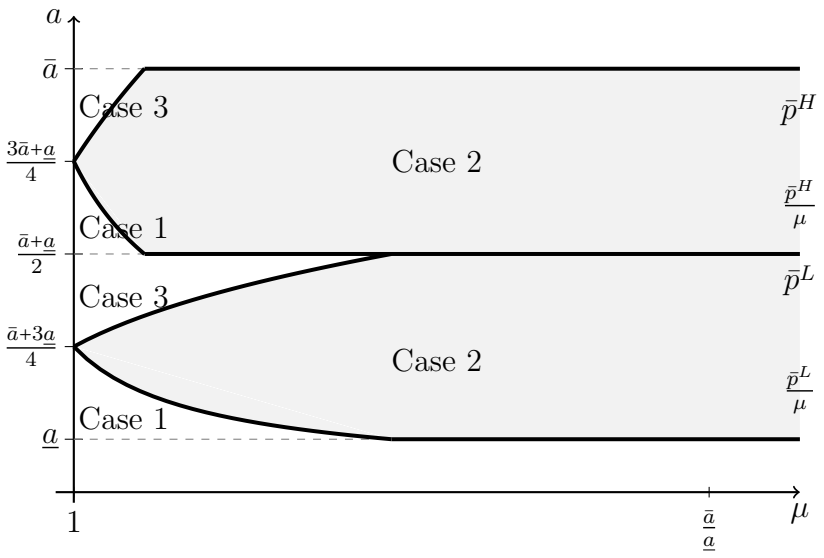


Figure 2.6: Case Thresholds for Students of Abilities $[\underline{a}, \bar{a}]$ under Ability Tracking

to perform higher. For low-ability students $\left(\frac{\bar{a}+3\underline{a}}{4}, \frac{\bar{a}+\underline{a}}{2}\right]$ the reverse is true, since they are now above the average. Especially these relatively high-ability types in the low track are the losers in terms of performance in an ability tracked system, whereas relatively low-ability students in the high track gain.

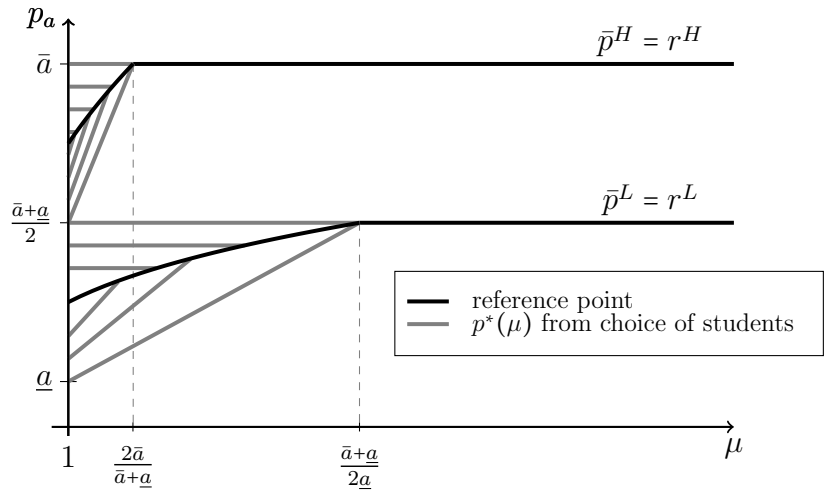


Figure 2.7: Equilibrium Performances as Functions of μ for a Discrete Choice of Students in a Non-Competitive Culture under Ability Tracking

2.4.3 Comprehensive Schooling vs. Ability Tracking

Formally comparing the average performances, $\bar{p}^C$ and $\bar{p}^T$, and the variances of performances, σ^C and σ^T, under the two regimes in non-competitive cultures leads to Proposi-

tion 6.

Proposition 6.

(i) Average performance in the Pareto dominant Nash equilibrium in a non-competitive culture is strictly higher under ability tracking than under comprehensive schooling if and only if $1 < \mu < \frac{\bar{a}^2+6\bar{a}\underline{a}+\underline{a}^2-(\bar{a}-\underline{a})\sqrt{\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2}}{8\underline{a}^2} \equiv \mu^*$.

(ii) The variance of performances in the Pareto dominant Nash equilibrium in a non-competitive culture is strictly higher under ability tracking than under comprehensive schooling for any $\mu > 1$.

A formal proof of Proposition 6 is given in Appendix 2.B. The first part of the Proposition states that ability tracking yields a strictly higher average performance for values of μ below a critical threshold. We know that for this critical μ^* it holds that $\frac{2\bar{a}}{\bar{a}+\underline{a}} < \mu^* < \frac{\bar{a}+\underline{a}}{2\underline{a}}$ (see Appendix 2.B), such that comprehensive schooling only yields a higher average performance for μ's high enough to have evoked at least a symmetric equilibrium in the high track. Since we are in a non-competitive culture and classes, where all students perform at the same high level are hardly observed, we argue that realistic levels of μ are well below this critical threshold. We conclude that ability tracking outperforms comprehensive schooling in non-competitive cultures in terms of average performance. The reason for this result is that motivating high-ability students is more beneficial than motivating low-ability students. We, however, face a trade-off between maximizing average performance and minimizing inequality, since a segregation of abilities leads to a higher variance of performances. This result mirrors past empirical research (e.g. Argys et al., 1996; Hoffer, 1992) that already argues that ability tracking benefits good students but harms low-ability students. A clear recommendation on the optimal school system cannot be given, but depends on the political objective.

Table 2.1 summarizes the results from comparing ability tracking and comprehensive schooling in competitive and non-competitive cultures.

Table 2.1: Summary of Results

	Average Performance $\bar{p}^C > \bar{p}^T$	Variance $\sigma^T > \sigma^C$
Non-Competitive Culture	$\mu > \mu^*$	$\mu > 1$
Competitive Culture	$\mu > 1$	$\mu > 1$

2.5 Discussion

The simplified linear value function used in our model might be subject to criticism. While our linear value function induces a constant marginal utility from performance below and above the reference point, a non-linear function allows for varying marginal utility depending on how close to the reference point a student's performance is. The value function from the original KT (1979) model is convex below the reference point and concave above, thereby modeling *diminishing sensitivity* with respect to the reference point. In our performance context this feature might be a more realistic mapping of student preferences. For instance, Gill and Prowse (2012) show in a real effort experiment on disappointment aversion that competing students are discouraged by a high effort choice of their opponent. Still, even though diminishing sensitivity is not modeled with our linear value function, the described effect is incorporated in the model by the convex cost function that depends negatively on students' ability. This evokes that students with lower ability are less motivated to reach the reference point since a marginal increase in performance is much more costly to them compared with high-ability students. Despite the intuitive importance of diminishing sensitivity, using a non-linear value function would lead to unrealistic results, since there are no interior solutions below the reference point. Another possible criticism is that symmetric equilibria at the reference point do only exist in our model, because of the kink in the linear value function at the reference point. In the case of a non-linear value function this corner solution would not exist, since the function is continuous. However, there would still be a convergence towards the reference point with increasing competitiveness, since the function becomes steeper around the reference point with increasing λ and increasing s.

So far we have only tackled the alternatives of one comprehensive school class versus a two track system. What happens if we institute more than two tracks? In the case of a competitive culture, the introduction of tracks is obviously never beneficial. Average performance decreases with the number of tracks and the variance of performances increases. In a non-competitive culture the picture is not as clear. Three equally sized tracks can yield a higher average performance than a two track system for μ sufficiently small. This critical μ is smaller than in Proposition 6. Comparing a four-track system with a three-track system the critical μ is even smaller. It thus depends on the exact preferences of the student body whether more than two tracks can still increase average performance. The trade-off between increasing average performance and minimizing inequality would, however, also aggravate.

Another question arising is whether it can be beneficial to introduce tracks in different

sizes. As an example consider Germany, where the school system is split into three schools for different ability types. The existing system has favored high inequality of performances (Baumert et al., 2001). Voices have become loud that call for an abolition of the stratification in favor of a comprehensive school system. However, many do not want to give up the "Gymnasium", the academic track in the school system, for fear of low performing students slowing down high performers. There are suggestions to merge the two lower school types, while keeping the "Gymnasium". Assuming that Germany is a non-competitive culture, take a look at Fig. 2.8 to see the impacts of such a policy in our model. Compared with a three track system the suggested system will decrease average performance for μ low enough.[4] The targets aimed at, however, are reached: For high performing students nothing changes and the variance of performances decreases. The losers are average ability students that have been in the middle track under the three track system. Low-ability students benefit from the merging of the two lower tracks, but this does not compensate the loss in performance of the average ability types.

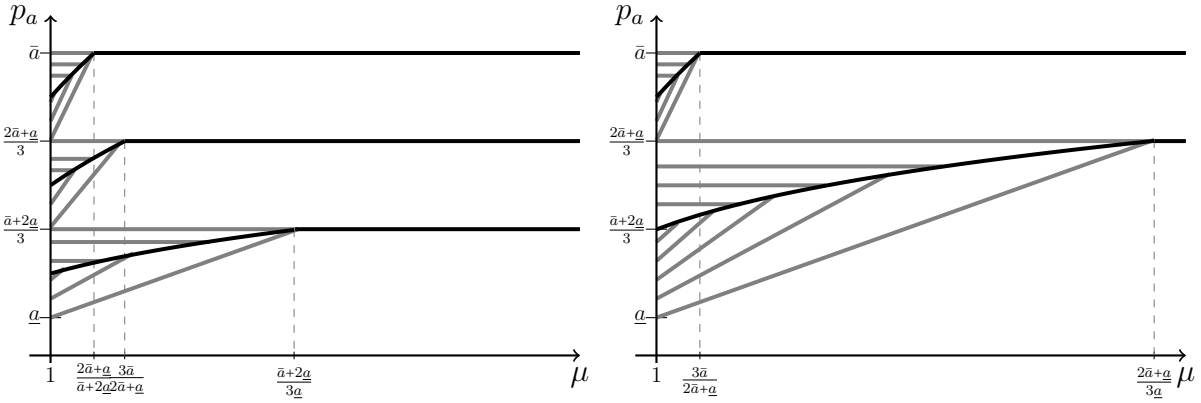


Figure 2.8: Equilibrium Performances for a Discrete Choice of Students under Three Equally Sized Tracks (Left) and after Merging Low and Middle Track (Right)

In the case of a competitive culture our model rests on the assumption that *all* students are motivated by loss aversion to reach the performance of the best student in class. Loss aversion with respect to a very high reference point, however, also means that students suffer a high loss of utility. Some might argue that students would rather opt out of competition in order to avoid this loss. For instance, Oettingen (1995) compares self-efficacy beliefs of students in competitive West Berlin and non-competitive East Berlin. She finds that the constant comparison with high achieving classmates can undermine the motivation of low performing students, that they have lower aspirations and give up

[4]Formally comparing the average performance of the two systems shows that a three track system yields a strictly higher average performance up to a critical μ^{**} that lies between the thresholds $\frac{3\bar{a}}{2\bar{a}+\underline{a}} < \mu^{**} < \frac{\bar{a}+2\underline{a}}{3\underline{a}}$, i.e. a higher μ than needed for the high and middle track to reach a symmetric equilibrium.

more readily in the face of difficulties. To take account of these effects, we can introduce a participation constraint, that states that students choose not to participate in competition as soon as their utility becomes non-positive, i.e. $p_a = 0$ if $u_a \leq 0$. Rearranging $u_a \leq 0$ we get another case in our best response function (2.3), with students performing zero if $a \leq \frac{2\lambda s r_a}{(1-s+\lambda s)^2}$. Concentrating on a competitive culture, Fig. 2.9 shows equilibrium performances as functions of the *degree of social comparison*, s, while λ is fixed on a level of 2.[5]

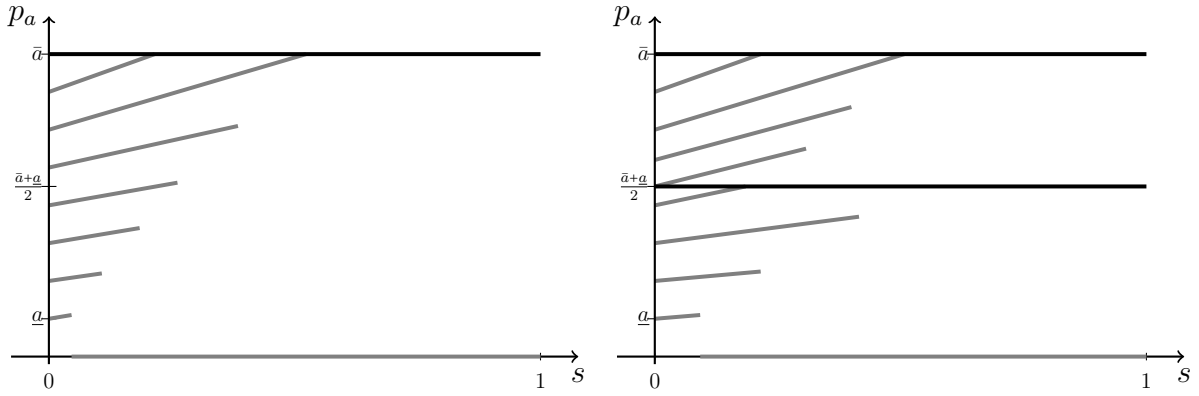


Figure 2.9: Equilibrium Performances Depending on the Degree of Social Comparison for a Discrete Choice of Students with Participation Constraint under Comprehensive Schooling (Left) and under Ability Tracking (Right)

From Fig. 2.9 it can be seen that the participation constraint leads to low-ability students opting out as s becomes larger. Under comprehensive schooling students drop out for lower levels of s than in the tracks.[6] Formally comparing average performance under the two regimes leads to the result that ability tracking can yield a higher average performance than comprehensive schooling as soon as students under comprehensive schooling start dropping out.[7] This is an important difference to the result without participation constraint. The mechanism for this becomes clear from Fig. 2.9. Under comprehensive schooling low-ability students suffer a huge utility loss, since the difference between their performance and the reference point is very high. Under ability tracking, however, low-ability students are not that far away from their reference point and are thus not as easily demotivated. Following this line of argument it would be beneficial to institute even more than two tracks.

[5]Measuring the coefficient of loss aversion has led to the common belief that it ranges around the number 2 (e.g. Johnson et al., 2006).

[6]Students under comprehensive schooling start dropping out as soon as $s > \frac{2\bar{a}-\underline{a}-\sqrt{4\bar{a}(\bar{a}-\underline{a})}}{\underline{a}}$ and students in the low track under ability tracking for $s > \frac{\bar{a}-\sqrt{\bar{a}^2-\underline{a}^2}}{\underline{a}}$. The more heterogeneous the ability distribution, i.e. the bigger the distance between $\bar{a}$ and $\underline{a}$, the earlier students drop out.

[7]This is the case, if the ability distribution is heterogeneous enough (i.e. if $\bar{a} > 1.3\underline{a}$), otherwise the critical s for ability tracking yielding a higher average performance is higher.

2.6 Conclusion

In this paper we integrate aspects of culture into a student effort-choice model with reference-dependent preferences. We show that a comprehensive school design is to be preferred in a competitive culture, whereas ability tracking yields a higher average performance in a non-competitive culture, but also leads to higher inequality. The difference in outcomes mainly stems from the difference in reference points chosen in the two extreme cultures. These results show that the cultural background of students matters for the decision on institutional design. A policy maker thus cannot simply rely on internationally identified best practice, and needs to take into account national culture.

The main factor that drives performances in our theory is loss aversion. Students are motivated because they strive to avoid a loss of utility due to a high difference between own performance and reference performance. A policy maker whose only aim is to increase performances can use this mechanism. Independent of whether a tracked or comprehensive school system is in place, the schooling context could be designed to reinforce competitive preferences. For instance, regular and frequent performance feedback and rankings could be provided to facilitate social comparisons. Teachers could highlight the best students' achievements in order to induce high reference points. However, students' utility or well-being is decreasing the further away the reference point and the fiercer the competition. In the worst case students might choose not to engage in competition.

There is much more research to be done in the area of culture's influence on students' learning behavior. Possible extensions to our model include the analysis of other possible (endogenous) reference points or the inclusion of uncertainty. In addition, the model could be tested empirically with field data and (or) in lab experiments.

Appendix

2.A Proof of Proposition 3

For the Proof of Proposition 3 we only consider the Pareto dominant Nash equilibrium, i.e. the best student in class performs at the lower bound of possible Nash equilibria.

(i) *Average performances:*
For students with ability $a \in (\frac{\bar{a}+\underline{a}}{2}, \bar{a}]$ (the high track students) performance is the same under comprehensive schooling and under ability tracking. For the remaining students $a \in [\underline{a}, \frac{\bar{a}+\underline{a}}{2}]$ (the low track students) performance under ability tracking is never higher than under comprehensive schooling. Moreover, for any $\mu > 1$ there exists $\epsilon > 0$ such that students with ability $a \in (\frac{\bar{a}+\underline{a}}{2} - \epsilon, \frac{\bar{a}+\underline{a}}{2}]$ have a lower performance under ability tracking than under comprehensive schooling. Hence, average performance must be strictly lower under ability tracking for any $\mu > 1$. **QED**

(ii) *Variance of performances*:
Part I:
Consider a random variable $X(\omega)$ (think of X as student's performance and of ω as student's ability). Now, let us construct a new variable $Y(\omega)$ such that $Y(\omega) = X(\omega)$ everywhere except on a certain positive-measure set Ω_c, on which $Y(\omega) \neq X(\omega)$. If, for every $\omega \in \Omega_c$ we have $|Y(\omega) - EX| > |X(\omega) - EX|$ then $Var(Y) > Var(X)$. In other words, if we create a new variable by moving away from the expected value of the old variable, then the variance of the new variable is higher than the variance of the old variable.
To see why this is true recall that the expected value minimizes the second moment, i.e. $E(X - EX)^2 \leq E(X - m)^2$ for any m. It follows that:

$$\begin{aligned} Var(Y) &= \int_\Omega (Y-EY)^2\,\mathrm{d}\omega = \int_{\Omega\setminus\Omega_c} (Y-EY)^2\,\mathrm{d}\omega + \int_{\Omega_c} (Y-EY)^2\,\mathrm{d}\omega \\ &\geq \int_{\Omega\setminus\Omega_c} (Y-EX)^2\,\mathrm{d}\omega + \int_{\Omega_c} (Y-EX)^2\,\mathrm{d}\omega \\ &> \int_{\Omega\setminus\Omega_c} (Y-EX)^2\,\mathrm{d}\omega + \int_{\Omega_c} (X-EX)^2\,\mathrm{d}\omega \\ &= \int_{\Omega\setminus\Omega_c} (X-EX)^2\,\mathrm{d}\omega + \int_{\Omega_c} (X-EX)^2\,\mathrm{d}\omega \\ &= \int_\Omega (X-EX)^2\,\mathrm{d}\omega = Var(X) \end{aligned}$$

Hence, $Var(Y) > Var(X)$.

Part II:
In this part we prove that the performance under ability tracking as a random variable can be obtained from the performance under comprehensive schooling in a way described in Part I. Therefore, $Var(p^T) = Var(Y) > Var(X) = Var(p^C)$.

Case 1: $\mu > \frac{2\bar{a}}{\bar{a}+\underline{a}}$ (*The high track students have reached a symmetric equilibrium*).

- Students $a \in (\frac{\bar{a}+\underline{a}}{2}, \bar{a}]$ (the high track students) have the same performance under ability tracking and under comprehensive schooling.
- Students $a \in [\underline{a}, \frac{\bar{a}+\underline{a}}{2}]$ (the low track students) can be divided into three groups, some of which may be empty. First group is the group of students whose performance is the same under ability tracking and comprehensive schooling. Second group is the group of students whose performance is at or below $\bar{p}^C$ under comprehensive schooling and the third group is the group of students whose performance is above $\bar{p}^C$ under comprehensive schooling. Notice that $\bar{p}^C$, given by $\bar{p}^C = \frac{1}{\bar{a}-\underline{a}}\left(\int_{\underline{a}}^{\frac{\bar{a}}{\mu}} \mu a\,\mathrm{d}a + \int_{\frac{\bar{a}}{\mu}}^{\bar{a}} \bar{a}\,\mathrm{d}a\right) = \frac{2\bar{a}^2\mu - \bar{a}^2 - \underline{a}^2\mu}{2\mu(\bar{a}-\underline{a})}$, is bigger than $\frac{3\bar{a}+\underline{a}}{4}$ in this case. This can be shown by inserting $\mu = \frac{2\bar{a}}{\bar{a}+\underline{a}}$ into $\bar{p}^C$ and simplifying to $\bar{p}^C = \frac{\bar{a}(3\bar{a}+5\underline{a})}{4(\bar{a}+\underline{a})}$. Since $\bar{p}^C$ increases in μ and by showing that $\frac{\bar{a}(3\bar{a}+5\underline{a})}{4(\bar{a}+\underline{a})} > \frac{3\bar{a}+\underline{a}}{4} \Leftrightarrow \bar{a} > \underline{a}$ it has to be true. Looking at the second group, we see that it consists of students that perform at $\frac{\bar{a}+\underline{a}}{2}$ under ability tracking, but perform at μa under comprehensive schooling. Since $\bar{p}^C \geq \mu a > \frac{\bar{a}+\underline{a}}{2}$ for all students in this group, performance under comprehensive schooling must be closer to $\bar{p}^C$ than under ability tracking. In the third group are students that perform at μa or $\bar{a}$ under comprehensive schooling, but at $\frac{\bar{a}+\underline{a}}{2}$ under ability tracking. Under ability tracking these students are at least $\frac{1}{4}(\bar{a}-\underline{a})$ away from $\bar{p}^C$, since $\bar{p}^C > \frac{3\bar{a}+\underline{a}}{4}$, while students under comprehensive schooling are at most $\frac{1}{4}(\bar{a}-\underline{a})$ away from $\bar{p}^C$. If

this group is non-empty, at most a measure-zero set of students is as far away from $\bar{p}^C$ under ability tracking as under comprehensive schooling. Since the second and the third group cannot be both measure-zero at the same time, the premises of Part I are satisfied and $Var(p^T) > Var(p^C)$.

Case 2: $1 < \mu < \frac{2\bar{a}}{\bar{a}+\underline{a}}$ (*There are no symmetric equilibria under ability tracking and under comprehensive schooling*)

- Students $a \in (\frac{\bar{a}+\underline{a}}{2}, \bar{a}]$ (the high track students) have the same performance under ability tracking and under comprehensive schooling.
- Students $a \in [\underline{a}, \frac{\bar{a}+\underline{a}}{2}]$ (the low track students) can be divided into 3 groups. The first group is a positive-measure group of students for whom $\mu a \leq \frac{\bar{a}+\underline{a}}{2}$, such that they have the same performance under comprehensive schooling and ability tracking. In the second group are those who perform at or below $\bar{p}^C$ under comprehensive schooling, and in the third group are those who perform above $\bar{p}^C$ under comprehensive schooling. Under comprehensive schooling all students perform at μa (the student with ability $\frac{\bar{a}+\underline{a}}{2}$ reaches $\bar{a}$ exactly for $\mu = \frac{2\bar{a}}{\bar{a}+\underline{a}}$), while all students in the second and third group perform at $\frac{\bar{a}+\underline{a}}{2}$ under ability tracking. For the second group we have $\bar{p}^C \geq \mu a > \frac{\bar{a}+\underline{a}}{2}$, such that performance under comprehensive schooling must be closer to $\bar{p}^C$ than under ability tracking. For the thrid group, notice that the highest-performing student in the this group has a performance of $\mu\frac{\bar{a}+\underline{a}}{2}$ under comprehensive schooling. Therefore, if $\mu\frac{\bar{a}+\underline{a}}{2} - \bar{p}^C < \bar{p}^C - \frac{\bar{a}+\underline{a}}{2}$, then all students in this group are further away from $\bar{p}^C$ under ability tracking than under comprehensive schooling. This inequality is equivalent to $\frac{\bar{a}+\underline{a}}{4}(\mu+1) < \bar{p}^C$, where $\bar{p}^C = \frac{2\bar{a}^2\mu - \bar{a}^2 - \underline{a}^2\mu}{2\mu(\bar{a}-\underline{a})}$. Given $\mu > 0$ this can be simplified to $(\bar{a}^2 - \underline{a})\mu < 2\bar{a}^2$. Using $\mu < \frac{2\bar{a}}{\bar{a}+\underline{a}}$ we can show that $(\bar{a}^2 - \underline{a}^2)\mu < (\bar{a}^2 - \underline{a}^2)\frac{2\bar{a}}{\bar{a}+\underline{a}} = 2\bar{a}\frac{(\bar{a}+\underline{a})(\bar{a}-\underline{a})}{\bar{a}+\underline{a}} = 2\bar{a}^2 - 2\underline{a}^2 < 2\bar{a}^2$. Hence, performance of all students either does not change, or is further away from $\bar{p}^C$ under ability tracking than under comprehensive schooling. As a result, we can apply Part I of the proof and it follows that $Var(p^T) > Var(p^C)$.

QED

2.B Proof of Proposition 6

For the Proof of Proposition 6 we only consider the Pareto dominant Nash equilibrium, i.e. if a symmetric equilibrium is reached, the students perform at the lower bound of the

possible Nash equilibria.

(i) *Average performance*:

Case 1: $\mu > \frac{\bar{a}}{\underline{a}}$ (*All students under comprehensive schooling and ability tracking have reached a symmetric equilibrium*)

It follows immediately that comprehensive schooling yields a strictly higher average performance, since average performance under comprehensive schooling is at $\bar{a}$ and under ability tracking at $\frac{1}{2}\left(\bar{a} + \frac{\bar{a}+\underline{a}}{2}\right)$, which is strictly lower.

Case 2: $\frac{\bar{a}+\underline{a}}{2\underline{a}} < \mu \leq \frac{\bar{a}}{\underline{a}}$ (*Under ability tracking all students have reached a symmetric equilibrium*)

We face a trade-off, since high track students gain from ability tracking, and low track students lose from ability tracking. We thus need to solve the following inequality to see whether the gain outweighs the loss:

$$\begin{aligned}
& \bar{p}^C > \tfrac{1}{2}\bar{p}^L + \tfrac{1}{2}\bar{p}^H \\
\Leftrightarrow \quad & \tfrac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{\sqrt{\mu}+1} > \tfrac{1}{2}\tfrac{\bar{a}+\underline{a}}{2} + \tfrac{1}{2}\bar{a} \\
\Leftrightarrow \quad & \tfrac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{\sqrt{\mu}+1} > \tfrac{3\bar{a}+\underline{a}}{4} \\
\Leftrightarrow \quad & 4(\bar{a}\sqrt{\mu} + \underline{a}\mu) > (3\bar{a} + \underline{a})(\sqrt{\mu} + 1) \\
\Leftrightarrow \quad & \bar{a}\sqrt{\mu} + 4\underline{a}\mu > \underline{a}\sqrt{\mu} + 3\bar{a} + \underline{a} \\
\Leftrightarrow \quad & (\bar{a} - \underline{a})\sqrt{\mu} + 4\underline{a}\mu > 3\bar{a} + \underline{a}
\end{aligned}$$

Since in this case $\mu > \frac{\bar{a}+\underline{a}}{2\underline{a}} \Leftrightarrow 4\underline{a}\mu > 2\bar{a}+2\underline{a}$, we have $(\bar{a}-\underline{a})\sqrt{\mu}+4\underline{a}\mu > (\bar{a}-\underline{a})\sqrt{\mu}+2\bar{a}+2\underline{a}$. Hence $(\bar{a} - \underline{a})\sqrt{\mu} + 4\underline{a}\mu > 3\bar{a} + \underline{a} \Leftarrow (\bar{a} - \underline{a})\sqrt{\mu} + 2\bar{a} + 2\underline{a} > 3\bar{a} + \underline{a} \Leftrightarrow (\bar{a} - \underline{a})\sqrt{\mu} > \bar{a} - \underline{a}$ which is always true for $\mu > 1$ and $\bar{a} > \underline{a}$. Therefore $\frac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{\sqrt{\mu}+1} > \frac{1}{2}\frac{\bar{a}+\underline{a}}{2} + \frac{1}{2}\bar{a}$ is also always true in this case.

Case 3: $\frac{2\bar{a}}{\bar{a}+\underline{a}} < \mu \leq \frac{\bar{a}+\underline{a}}{2\underline{a}}$ (*The high track students under ability tracking have reached a symmetric equilibrium*)

We again face a trade-off so that we need to solve the following inequality:

$$\begin{aligned}
& \bar{p}^C > \tfrac{1}{2}\bar{p}^L + \tfrac{1}{2}\bar{p}^H \\
\Leftrightarrow \quad & \tfrac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{1+\sqrt{\mu}} > \tfrac{1}{2}\tfrac{(\bar{a}+\underline{a})\sqrt{\mu}+2\underline{a}\mu}{2(1+\sqrt{\mu})} + \tfrac{1}{2}\bar{a} \\
\Leftrightarrow \quad & \tfrac{1}{2}(\bar{a} - \underline{a})\sqrt{\mu} + \underline{a}\mu > \bar{a} \\
\Leftrightarrow \quad & \tfrac{1}{4}(\bar{a} - \underline{a})^2\mu > (\bar{a} - \underline{a}\mu)^2 \\
\Leftrightarrow \quad & -\underline{a}^2\mu^2 + \left(\tfrac{1}{4}\bar{a}^2 + \tfrac{3}{2}\bar{a}\underline{a} + \tfrac{1}{4}\underline{a}^2\right)\mu > \bar{a}^2 \\
\Leftrightarrow \quad & \left(\mu - \tfrac{\bar{a}^2+6\bar{a}\underline{a}+\underline{a}^2}{8\underline{a}^2}\right)^2 < \tfrac{(\bar{a}-\underline{a})^2(\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2)}{64\underline{a}^4} \\
\Leftrightarrow \quad & \mu > \tfrac{\bar{a}^2+6\bar{a}\underline{a}+\underline{a}^2-(\bar{a}-\underline{a})\sqrt{\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2}}{8\underline{a}^2} \equiv \mu^*
\end{aligned}$$

(the second solution of the quadratic function is not a solution of the original root function)

We can show that $\frac{2\bar{a}}{\bar{a}+\underline{a}} < \mu^* < \frac{\bar{a}+\underline{a}}{2\underline{a}}$, such that comprehensive schooling yields a strictly higher average performance for $\mu > \mu^*$ and ability tracking yields a strictly higher average performance for $\mu < \mu^*$:

$$
\begin{aligned}
& \mu^* < \tfrac{\bar{a}+\underline{a}}{2\underline{a}} \\
\Leftrightarrow \quad & \tfrac{\bar{a}^2+6\bar{a}\underline{a}+\underline{a}^2-(\bar{a}-\underline{a})\sqrt{\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2}}{8\underline{a}^2} < \tfrac{\bar{a}+\underline{a}}{2\underline{a}} \\
\Leftrightarrow \quad & \bar{a}^2 + 2\bar{a}\underline{a} - 3\underline{a}^2 - (\bar{a}-\underline{a})\sqrt{\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2} < 0 \\
\Leftrightarrow \quad & (\bar{a}-\underline{a})\left(\bar{a}+3\underline{a}-\sqrt{\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2}\right) < 0 \\
\Leftrightarrow \quad & (\bar{a}+3\underline{a})^2 < \bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2 \\
\Leftrightarrow \quad & \underline{a}^2 < \bar{a}\underline{a} \quad \text{(true for all } \bar{a} > \underline{a} > 0\text{)}
\end{aligned}
$$

$$
\begin{aligned}
& \mu^* > \tfrac{2\bar{a}}{\bar{a}+\underline{a}} \\
\Leftrightarrow \quad & \tfrac{\bar{a}^2+6\bar{a}\underline{a}+\underline{a}^2-(\bar{a}-\underline{a})\sqrt{\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2}}{8\underline{a}^2} > \tfrac{2\bar{a}}{\bar{a}+\underline{a}} \\
\Leftrightarrow \quad & (\bar{a}+\underline{a})\left(\bar{a}^2+6\bar{a}\underline{a}+\underline{a}^2\right) - 16\bar{a}\underline{a}^2 - (\bar{a}-\underline{a})(\bar{a}+\underline{a})\sqrt{\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2} > 0 \\
\Leftrightarrow \quad & \tfrac{\bar{a}^2+8\bar{a}\underline{a}-\underline{a}^2}{\bar{a}+\underline{a}} > \sqrt{\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2} \\
\Leftrightarrow \quad & \tfrac{\left(\bar{a}^2+8\bar{a}\underline{a}-\underline{a}^2\right)^2}{(\bar{a}+\underline{a})^2} > \bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2 \\
\Leftrightarrow \quad & \left(\bar{a}^2+8\bar{a}\underline{a}-\underline{a}^2\right)^2 - \left(\bar{a}^2+14\bar{a}\underline{a}+\underline{a}^2\right)(\bar{a}+\underline{a})^2 > 0 \\
\Leftrightarrow \quad & 32\bar{a}^2\underline{a}^2 - 28\bar{a}\underline{a}^2 > 0 \quad \text{(true for all } \bar{a} > \underline{a} > 0\text{)}
\end{aligned}
$$

Case 4: $1 < \mu \leq \frac{2\bar{a}}{\bar{a}+\underline{a}}$ (*There are no symmetric equilibria under ability tracking and under comprehensive schooling*)
We again face a trade-off so that we need to calculate:

$$
\begin{aligned}
& \bar{p}^C > \bar{p}^T \\
\Leftrightarrow \quad & \bar{p}^C > \tfrac{1}{2}\bar{p}^L + \tfrac{1}{2}\bar{p}^H \\
\Leftrightarrow \quad & \tfrac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{1+\sqrt{\mu}} > 0.5\tfrac{(\bar{a}+\underline{a})\sqrt{\mu}+2\underline{a}\mu}{2(1+\sqrt{\mu})} + 0.5\tfrac{(\bar{a}+\underline{a})\mu+2\bar{a}\sqrt{\mu}}{2(1+\sqrt{\mu})} \\
\Leftrightarrow \quad & 4\bar{a}\sqrt{\mu} + 4\underline{a}\mu > (\bar{a}+\underline{a})\sqrt{\mu} + 2\underline{a}\mu + (\bar{a}+\underline{a})\mu + 2\bar{a}\sqrt{\mu} \\
\Leftrightarrow \quad & (\bar{a}-\underline{a})\sqrt{\mu} > (\bar{a}-\underline{a})\mu \\
\Leftrightarrow \quad & \sqrt{\mu} > \mu
\end{aligned}
$$

Since $\mu > 1$ we end in a contradiction, such that we can conclude that ability tracking yields a strictly higher average performance in this case.
QED

(ii) *Variance of performances*:

Case 1: $\mu \geq \frac{\bar{a}}{\underline{a}}$ (*All students under comprehensive schooling and ability tracking have reached a symmetric equilibrium*)
Notice that $\bar{p}^C$ takes only one value and $\bar{p}^T$ takes equally likely two values, hence $Var(p^T) > Var(p^C) = 0$.

For the other cases we prove that the performance under ability tracking as a random variable can be obtained from the performance under comprehensive schooling in a way described in Part I of the Proof of Proposition 3 (ii).

Case 2: $\frac{\bar{a}+\underline{a}}{2\underline{a}} \leq \mu < \frac{\bar{a}}{\underline{a}}$ (*Under ability tracking all students have reached a symmetric equilibrium*)

- Students $a \in (\frac{\bar{a}+\underline{a}}{2}, \bar{a}]$ (the high track students) perform at $\bar{p}^C$ or higher at a under comprehensive schooling. Under ability tracking all students perform at $\bar{a}$. Performance a for the considered students is always closer to $\bar{p}^C$ and only in the case of student $\bar{a}$ as far away from $\bar{p}^C$ under comprehensive schooling as under ability tracking.
- Students $a \in [\underline{a}, \frac{\bar{a}+\underline{a}}{2}]$ (the low track students) perform at $\bar{p}^C$ or lower at μa under comprehensive schooling. They all perform greater or equal to $\frac{\bar{a}+\underline{a}}{2}$, since the lowest ability student $\underline{a}$ performs at $\mu\underline{a}$, which is in this case at least $\frac{\bar{a}+\underline{a}}{2\underline{a}}\underline{a} = \frac{\bar{a}+\underline{a}}{2}$. Under ability tracking all students perform at $\frac{\bar{a}+\underline{a}}{2}$. Since $\bar{p}^C > \frac{\bar{a}+\underline{a}}{2}$ (it is easy to show, for $\mu > 1$: $\bar{p}^C = \frac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{\sqrt{\mu}+1} > \frac{\bar{a}\sqrt{\mu}+\underline{a}}{\sqrt{\mu}+1} > \frac{\bar{a}+\underline{a}}{2}$ where the last inequality comes from the fact that on the left-hand side we are taking a weighted average with a higher weight on a larger element) all students under comprehensive schooling are closer and only in the case of student $\underline{a}$ at least as close to $\bar{p}^C$ as students under ability tracking.
- Therefore, $Var(p^T) > Var(p^C)$ in this case.

Case 3: $\frac{2\bar{a}}{\bar{a}+\underline{a}} < \mu \leq \frac{\bar{a}+\underline{a}}{2\underline{a}}$ (*The high track students under ability tracking have reached a symmetric equilibrium*)

- For students $a \in (\frac{\bar{a}+\underline{a}}{2}, \bar{a}]$ (the high track students) the same as in case 2 applies.
- Students $a \in [\underline{a}, \frac{\bar{a}+\underline{a}}{2}]$ (the low track students) perform at $\bar{p}^C$ or below at μa under comprehensive schooling. The students that perform at μa under com-

prehensive schooling cannot perform closer to $\bar{p}^C$ under ability tracking, since it is not possible to perform higher than μa according to the best response function. Thus, there are no students who perform closer to $\bar{p}^C$ under ability tracking. On the other hand, those students who perform at $\bar{p}^C$ under comprehensive schooling, perform further away from $\bar{p}^C$ under ability tracking, since they perform smaller or equal to $\frac{\bar{a}+\underline{a}}{2}$, but $\bar{p}^C > \frac{\bar{a}+\underline{a}}{2}$. Hence, there is a positive-measure group of students whose performance under ability tracking is further away from $\bar{p}^C$ than under comprehensive schooling, while no student gets closer to $\bar{p}^C$.

- Therefore, $Var(p^T) > Var(p^C)$ in this case.

Case 4: $1 < \mu \leq \frac{2\bar{a}}{\bar{a}+\underline{a}}$ (*There are no symmetric equilibria under ability tracking and under comprehensive schooling*)

- Students $a \in (\frac{\bar{a}+\underline{a}}{2}, \bar{a}]$ (the high track students) perform at $\bar{p}^C$ or above at a under comprehensive schooling. Under ability tracking all students perform above $\bar{p}^C$, since performance of the lowest ability student is at least $\mu\frac{\bar{a}+\underline{a}}{2}$ which is bigger than $\bar{p}^C$ (it is easy to see that $\mu\frac{\bar{a}+\underline{a}}{2} > \frac{\bar{a}\sqrt{\mu}+\underline{a}\mu}{\sqrt{\mu}+1}$ must be true, since $\sqrt{\mu}+1 > 2$ and $\bar{a}\mu + \underline{a}\mu > \bar{a}\sqrt{\mu} + \underline{a}\mu$ for $\mu > 1$.). Hence all students that perform at $\bar{p}^C$ under comprehensive schooling perform further away under ability tracking. The students that perform at a under comprehensive schooling do not perform closer to $\bar{p}^C$ under ability tracking, since all other possible performances at μa or $\bar{p}^H$ are even bigger than a ($\bar{p}^H > a$ must be true, since according to the best response function $\bar{p}^H$ is only played by students with $a < \bar{p}^H$) and thus further away from $\bar{p}^C$. Hence, no student gets closer to $\bar{p}^C$ under ability tracking, while there is a positive-measure group of students whose performance under ability tracking is further away from $\bar{p}^C$.
- For students $a \in [\underline{a}, \frac{\bar{a}+\underline{a}}{2}]$ (the low track students) the same as in case 3 applies.
- Therefore, $Var(p^T) > Var(p^C)$ in this case.

QED

Chapter 3

Does the Impact of Ability Grouping vary with the Culture of Competitiveness? - Evidence from PISA 2012

Abstract

In this paper theoretical hypotheses from Thiemann (2017) are tested for their empirical relevance. According to theory comprehensive schooling and ability grouping yield different results in terms of average student performance in countries that differ in their culture of competitiveness. The predictions are tested using a country-level indicator on the appraisal of competition from the World Values Survey. Educational achievement data is from PISA 2012, covering 34 countries and more than 10,000 schools of which data on the school's policy of ability grouping is available. To overcome possible endogeneity of ability grouping an instrumental variable approach is employed, using the number of schools a school regionally competes with as an instrument. The estimation shows that ability grouping in some or all classes increases average student achievement in competitive cultures and decreases average student achievement in non-competitive cultures.

JEL-Code: I20, I24, O15, H75

Keywords: Ability Grouping, Ability Tracking, Culture, Competitiveness, PISA, Education Production Function, Instrumental Variables, Quantile Regression

3.1 Introduction

Among the top performers of the most recent PISA (Programme for International Student Assessment) study 2012 are countries like Switzerland, the Netherlands or Singapore (OECD, 2013b), all countries that rigidly sort students into different schools based on their ability. Still, there are also countries like Finland and Japan at the top of the ranking, where students of very heterogeneous abilities are all taught together in one class. This suggests that different approaches are successful in different countries. In a recently published report "The learning curve" (The Economist Intelligence Unit, 2012) about the search for international best practices in education, the authors admit that none were found. They describe the way in which differences in the country-specific learning process transform inputs into outputs as a "black box" which is difficult to predict or quantify consistently. A possible reason for this finding is that countries differ in their cultures of teaching and learning. The question focused on in this paper is whether learning in small ability segregated groups (ability grouping) is to be preferred over learning in a class with students of heterogeneous abilities and backgrounds (comprehensive schooling) and to what degree the answer depends on student characteristics that vary with culture. In particular we focus on the country-specific culture of competitiveness that might influence the effect of AG on student achievement. Competitiveness thereby refers to the innate drive and desire of students to socially compare and outperform peers.

Theoretical predictions on this topic are formulated by Thiemann (2017). Here a model of student decision making is developed that explains the different effect of AG by peer effects that have different mechanisms in competitive and non-competitive cultures. Competitive cultures are defined as cultures where social comparison is an important part of the student's utility function. More precisely, competitive students are assumed to compare their own performance with the *best* performance in class, which serves as a reference point in the reference-dependent utility function. In addition, competitive students are assumed to suffer a lot from failures in school, which translates into a high loss aversion. The opposite is true for non-competitive cultures, where social comparison does only weakly influence the students' effort choice and where the *average* performance in class is the reference point of comparison. These assumptions are built on the description of culturally different learning styles by the cross-cultural researcher Hofstede (1986). The hypotheses derived from this model are taken to the data of PISA 2012 in this paper. The aim is to find empirical evidence for the following theoretical predictions. First, we seek general evidence for the existence of an influence of culture on the effect of AG on student performance. Second, predictions on the performance of students in

competitive cultures can be derived from the model. In the simple case with linear utility from Thiemann (2017) comprehensive schooling yields a higher average performance than AG in competitive cultures. In these cultures students have high reference points, such that comprehensive schooling provides all students with the motivating force of a high reference for comparison. In a system with AG, where high-ability students are sorted into a high track and low-ability students into a low track, this positive effect is restricted to the students in the high track. Assuming non-linear utility functions, thereby modeling diminishing sensitivity with respect to the reference point, changes this result. This assumption takes into account the hypothesis that being just below the reference point induces a higher motivation than being further away. In comprehensive schools low-ability students would thus not experience much motivation if they compare with the best student whose performance is too high to be reached. Classes of rather homogenous abilities would then be preferred. This view is also supported by an extension to the model including a participation constraint in Thiemann (2017), which states that students choose not to participate in competition (do not perform at all), if their utility from optimal performance is negative. This extension takes into account that many students would opt out of competition in order to avoid a high loss of utility evoked by loss aversion with respect to a very high reference point. This problem is particularly relevant in comprehensive schools where classes consist of heterogeneous abilities. Competitive students with low ability easily give up in these classes, since the reference point is too far away. Finding evidence for AG being beneficial in competitive cultures would thus support the idea of diminishing sensitivity and participation constraints.

Third, there are predictions for students from non-competitive cultures. The linear model predicts AG to yield a higher average performance than comprehensive schooling in non-competitive cultures. Since students' reference point is the average performance in class, AG can be better at motivating high-ability students since their reference point is higher in a high track than under comprehensive schooling. This effect may on average outweigh the negative effect of AG for low-ability students. The impact of diminishing sensitivity and participation constraints would not change this result, since both assumptions work in general in favor of AG.

Fourthly, theory from Thiemann (2017) predicts that the overall variance of student achievement increases under AG. This is because AG is, both in competitive and non-competitive cultures, detrimental for low-ability students, but beneficial for high-ability students at least in the linear model. If we find evidence for higher variance under comprehensive schooling, this might be evidence for diminishing sensitivity or participation constraints.

Where the theory underlying this paper can describe social preferences and the mechanisms of peer effects in different cultures very precisely, the reality is much more complex. Preferences and likewise culture are not directly observable. Education can be viewed as a black box, where educational inputs (spending, class size, ability grouping) go in and culture-specific outputs are produced. This paper tries to open parts of this black box by using a survey question from the World Values Survey (WVS) (Inglehart, 2014) to derive a measure for country-level competitive preferences.

The theoretical predictions are tested by estimating a typical education production function. This function explains student achievement by multilevel variables: Student background and family information, school characteristics and country specific factors. The empirical estimation of this function uses PISA 2012 math data including roughly 250.000 student observations from 34 countries. The regressor of interest is a measure for AG, which is based on school principals' reports within the PISA study on whether the school groups math classes according to student ability. This school level variable on AG is interacted with the mentioned country level indicator for competitiveness from the WVS. In a least squares approach including country fixed effects the average effect of AG on performance, holding all other factors constant, is estimated. Furthermore, quantile regressions are performed to test the effect of AG across the conditional achievement distribution of students. This also yields insights on the effect of AG on the overall variance of student achievement. As a robustness check an instrumental variable (IV) approach is performed to control for possible endogeneity of the AG variable. This concern exists because of possible student self-selection into schools that perform a certain grouping policy.

The analysis of the PISA 2012 data shows, first and foremost, that culture *does* matter for the effect of AG on student performance. According to the estimation results show students in competitive cultures benefit from AG, whereas students in non-competitive cultures perform lower if they are grouped according to ability. This holds for all students along the conditional achievement distribution, only that students at the tails are generally less affected than those closer to the median. The effect of AG on the variance of achievement is not significantly different from zero in either culture. The IV approach proves to be unnecessary, since endogeneity of the AG variable can be rejected.

The remainder of this paper is organized as follows: Section 3.2 provides an overview of the related literature. In Section 3.3 the data used for the analysis is described in detail. In Section 3.4 the estimation method is outlined. Section 3.5 reports estimation results. Section 3.6 provides the IV approach and Section 3.7 further robustness checks. Section 3.8 concludes.

3.2 Related Literature

The question of how AG (sometimes also called *ability streaming* or *ability tracking*) affects students' performance has occupied researchers since the early 20th century. Especially in the USA and the United Kingdom economists have tried to estimate the effect of AG on performance using small student samples from grouped and ungrouped schools. An early literature review is provided by Slavin (1990). The evidence is very mixed, but mostly no strong effect of AG has been found. Since the 1990s bigger data sets are available which has given rise to new approaches in finding an effect of AG. A more recent literature review is provided by Meier and Schütz (2007). There are roughly three strands of literature that empirically analyze the effects of AG: First, there are many studies from the USA that exploit the variation of AG policies within and across American High Schools (e.g. Hoffer, 1992; Argys et al., 1996; Betts and Shkolnik, 2000). Second, there is a strand of literature that uses data from international achievement tests to analyze differences across countries that differ in their national tracking policies (e.g. Ammermüller, 2005; Hanushek and Woessmann, 2006). Third, some studies exist that exploit data from policy reforms and institutional changes in a country's school system (e.g. Pekkarinen et al., 2009; Galindo-Rueda and Vignoles, 2007)). The approach used in this paper combines the first two strands, since effects of AG at the school level are examined, while using international achievement data that includes a variety of countries. To the best of our knowledge there is no empirical literature on the effect of culture on outcomes in education in combination with the effect of AG.

The US studies that analyze AG policies across and within schools struggle with the problem of selection bias. The students' school choice and thus track placement might be affected by unobserved student characteristics such as innate ability, motivation or socio-economic factors. Researchers have developed different strategies to overcome this problem. Hoffer (1992) uses the Longitudinal Study of American Youth (LSAY) to examine the effect of AG on achievement growth from seventh to ninth grade. To overcome criticisms of selection bias Hoffer employs a propensity score approach. He runs a probit regression to predict the probability of high or low track placement for every student and then estimates the effect of actual group placement for different quintiles of these probability distributions. Hoffer does not find a significant effect of grouping on overall average achievement, but finds a moderate positive effect for students in the high group and a stronger negative effect for students in the low group.

Argys, Rees, and Brewer (1996) estimate a selection model to overcome the selection bias problem. They use the US National Education Longitudinal Survey to estimate the

effect of AG on the growth of students' math test scores from 8th to 10th grade. The first-stage of their approach is a multinomial logit model, where track placement for every student is predicted by usage of the following instruments: the racial ethnic make-up of the student body, the region in which the school is located and an indicator for whether the school is located in an urban, suburban or rural community. From these regressions they calculate selectivity correction terms (inverse Mills ratios). In a second stage they include these terms in education production functions that they estimate separately for every track (honors, academic, vocational). The predicted mean achievements are then compared to mean achievement in a heterogeneous class. They find that students in lower tracks would gain from de-tracking, while students in higher tracks would lose. Overall de-tracking would decrease average test scores by 2 %. The Argys et al. (1996) approach is criticized by Figlio and Page (2002), who remark that no evidence on the exogeneity of the instruments is provided.

Betts and Shkolnik (2000) control for unobserved innate ability and motivation by using information on the ability level of the class provided by the teacher. Achievement data is from the LSAY. They do not find an effect of AG on overall achievement, but find that low-ability students are not affected, middle ability students are harmed and high-ability students gain. As a robustness check they estimate a selection model comparable to Argys et al. (1996) using as instruments the percentage of black students in the school, the percentage of students who receive full federal lunch assistance and students' test score relative to the average for his or her grade.

Figlio and Page (2002) use the same data set as Argys et al. (1996) to determine the effect of AG on achievement growth from 8th to 10th grade. They divide the student achievement distribution from 8th grade into top, middle and bottom third and estimate separate regressions for each group. They include a dummy on whether the principal reported that the school applies AG, but find no significant effect in any subgroup. To overcome selection bias, they also estimate a two-stage-least-squares approach using as instruments: the number of schools in the region, the fraction of Reagan voters in the region and the number of academic courses required for state graduation. They only use the interactions of these variables as exogenous instruments to ensure that they are not correlated with achievement. Evidence from this approach suggests that AG has a positive effect on the bottom third and a slight negative effect on students in the top third.

Just like the estimations in this strand of literature, also our estimation might be affected by the problem of selection bias, since school level data on AG is used. In line with Figlio and Page (2002) an instrumental variable approach is employed, contributing to the literature by suggesting as instrument the number of schools that the given school

competes with. This strategy proves to be unnecessary since PISA data provides such a rich set of student background variables that renders the problem of unobserved student characteristics nonexistent.

The second strand of literature uses international achievement studies such as PISA, TIMMS (Trends in Mathematics and Science Study) or PIRLS (Progress in International Reading Literacy Study) to determine the effect of AG. These studies usually define AG on a country level, using different measures such as years spent in tracks, share of students in vocational tracks, the timing of tracking or simply a dummy that indicates whether the country has a grouping policy. Using country-level data comes with the problems of a lack of observations and the difficulty of controlling for all institutional and cultural differences between countries. Ammermüller (2005) tries to overcome this problem by estimating difference-in-difference effects using primary school data from PIRLS and secondary school data from PISA for 12 countries. He can thus cancel out all institutional and cultural country specific effects that do not change over schooling time. His focus is on the question of how changes in institutional variables such as AG influence the strength of the effect of family background variables on achievement. Measuring AG by the number of schools or tracks available to students in secondary schooling, he finds that this variable in combination with parents' education and origin has a positive effect on achievement.

Hanushek and Woessmann (2006) follow a similar approach in also estimating difference-in-difference effects, thus exploiting the fact that all countries that have an AG policy only start sorting after primary schooling. They regress secondary school test scores from TIMMS and PISA on matched primary school test scores from PIRLS and TIMSS. The matching of the tests produces different data sets with 18-26 countries depending on wave and subject. Including a dummy indicating whether the country has a tracking policy they find evidence of a weak negative effect of AG on average performance and a stronger positive effect on inequality, measured by the standard deviation of achievement and differences between percentiles.

Brunello and Checchi (2007) use data from different sources to measure the effect of family background, measured by parental education, in combination with AG on outcome variables for young adults such as earnings, employment, educational attainment and literacy. Their data set spans over several years and includes 12-25 countries depending on the outcome variable. To control for country specific effects they include country by cohort dummies. They find that the effect of family background becomes stronger with AG. Another result is that AG causes a stronger dispersion of earnings.

This paper can contribute to this literature by using school-level data, that has the advantage that it has a much higher variance in the AG variable and country fixed effects

can be included to control for unobserved country specific factors.

The third strand of literature investigates policy reforms to learn from institutional changes. There are two papers by Galindo-Rueda and Vignoles (2007) and Manning and Pischke (2006) that investigate the gradual change from a selective to a comprehensive school system in England and Wales in the 60s and 70s. Whereas Galindo-Rueda and Vignoles find that a selective school system favors high-ability students, Manning and Pischke (2006) do not find any significant effects. Pekkarinen et al. (2009) take a look at the Finnish reform from a two-track system to a comprehensive school system that took place gradually in 1972-1977. The major finding from their difference-in-difference approach is that the reform reduced inequality, as proven by a significant drop in the intergenerational income elasticity by 23%.

Overall, positive effects of AG are usually assigned to the channel of better targeted pedagogy (see Cortes and Goodman, 2014; Duflo et al., 2011), while the channel of peer effects is made responsible for the positive effects on high skilled students and negative effects on low skilled students (see Argys et al., 1996; Hoffer, 1992). The mixed evidence from past research has therefore several reasons. First, there is a lack of disentanglement of the channels through which the effects of AG work. Second, there are many different empirical approaches and different definitions of AG. Third, many of the reviewed studies are based on different subject pools from different countries and therefore different cultures. This paper contributes to the first point by aiming at explaining effects through the channel of peer effects only. Most importantly we also contribute to the third point by showing that effects of AG differ between cultures of competitiveness and thereby provide an explanation for the mixed evidence from past research.

3.3 Data

3.3.1 Student Achievement Data

Student achievement is measured using data from the 2012 PISA study. In the 2012 wave the acquired knowledge of about 510,000 15-year-old students from 65 countries is assessed in three key areas: reading, mathematics, science and problem solving. In the focus area, mathematics, students solved paper and pencil test questions that assess their capacity to formulate, employ and interpret mathematics in a variety of contexts. The test lasts about 2 hours and subsequently students have to fill in a background questionnaire. The individual student assessment is measured on a scale that is based on a mean for OECD countries of 500 points and a standard deviation of 100 points that were set in PISA

2003 when the first PISA scale was developed (OECD, 2013b). The sampling procedure of PISA is a two-stage sampling design. For each country first a sample of schools is selected from a complete list of schools containing the student population of interest. Then, a simple random sample of 35 students from the 15-year-old student population is drawn from within the selected schools (OECD, 2014). The principal of the selected school is asked to complete a questionnaire on school characteristics, generating a data set including student and school level variables.

The data used for this paper includes only 34 of the 65 PISA countries. The number of countries is reduced for two reasons. First, cultural data is not available for all countries.[1] Second, we only include countries that have a comprehensive school system on a national level. The variable used as an identifier for AG in this paper is a school level variable that yields information on whether classes within the school are grouped according to ability (see Section 3.3.3 for more details). This approach yields more variance and observations than comparing tracked and comprehensive school systems on a country level. Since PISA does not take into account that schools might be part of a nationally tracked school system, the variable on AG within schools might be biased. The school might already be a selection of low or high-ability students, if the whole school is part of a nationally tracked system. The effect of additional grouping on performance in these schools is not the same as in a comprehensive school system. To solve this problem countries with a tracked school system are excluded from the estimation.[2] Furthermore, we delete all observations of first or second generation immigrant students. Since we assume that competitiveness is a value that is transmitted from parents to their children, we cannot assign national culture to immigrants. This results in a loss of 27,157 observations. Table 3.1 lists the 34 countries, the number of schools and students included in the analysis. Even though the PISA sample is generally biased towards developed countries, the latest test from 2012 includes a wide variety of cultures, including non-OECD countries from South-America and Asia. Altogether 10,588 schools and 251,972 student observations are included.

The reasons for using PISA 2012 data are, first, that it provides a huge database covering a large number of countries to ensure that there is enough variation in terms of culture. Second, the PISA 2012 questionnaire contains a question on AG that fits the purposes of this study. Since the focus of the PISA 2012 study was on mathematics, the question on AG asks specifically for grouping in *mathematics* classes. This is ideal, since in

[1] These countries are: Costa Rica, Cyprus, Denmark, Greece, Ireland, Iceland, Israel, Liechtenstein, Macao-China, Montenegro, Portugal, Shanghai-China, Tunisia, United Arab Emirates.

[2] These countries are: Germany, The Netherlands, Belgium, Turkey, Austria, Switzerland, Luxembourg, Bulgaria, Romania, Hungary, The Czech Republic, Uruguay, Singapore, Korea, Italy, Croatia. Information on tracked school systems is taken from the OECD (2013a, p.78).

Table 3.1: Student and School Observations by Country

code	country	schools	students	code	country	schools	students
ALB	Albania	187	4,246	LTU	Lithuania	209	4,470
ARG	Argentina	220	5,437	LVA	Latvia	203	3,964
AUS	Australia	759	11,738	MEX	Mexico	1,453	33,050
BRA	Brazil	648	16,861	MYS	Malaysia	163	5,076
CAN	Canada	862	17,435	NOR	Norway	186	4,038
CHL	Chile	218	6,737	NZL	New Zealand	155	2,842
COL	Colombia	323	8,565	PER	Peru	238	5,993
ESP	Spain	862	22,338	POL	Poland	166	4,194
EST	Estonia	202	4,359	QAT	Qatar	143	4,718
FIN	Finland	301	7,433	RUS	Russia	224	4,614
FRA	France	200	3,528	SRB	Serbia	138	3,950
GBR	Great Britain	473	10,903	SWE	Sweden	207	4,033
HKG	Hong Kong	147	3,054	TAP	Taiwan	163	6,016
IDN	Indonesia	206	5,533	THA	Thailand	239	6,571
JOR	Jordan	225	5,960	TUN	Tunisia	150	4,303
JPN	Japan	190	6,293	USA	USA	155	3,881
KAZ	Kazakhstan	211	4,885	VNM	Vietnam	162	4,954
					Total	10,588	251,972

the analysis only the achievement data from the mathematics test is used. This is because math data is generally viewed as being most comparable across countries (Hanushek et al., 2013). Third, PISA data has the huge advantage that test scores are comparable across all students, schools and countries. This is important, since comparing school grades of grouped and ungrouped students might otherwise be biased because of different grading practices. Figure 3.1 shows 2012 mean performance in mathematics for the 34 countries in the sample. Mean performance is highest in East-Asian countries and lowest in South-American and South-East-Asian countries.

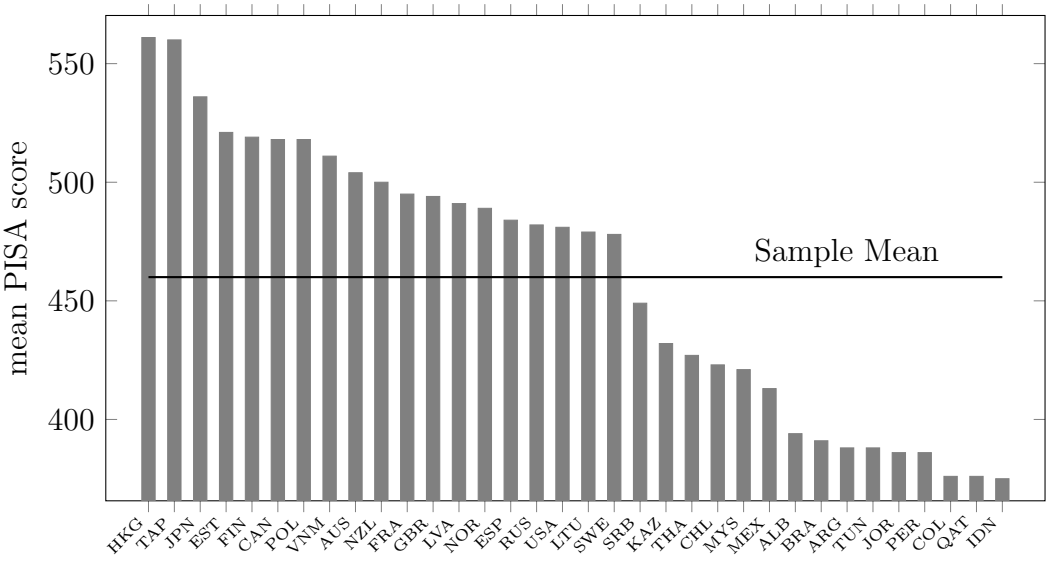


Figure 3.1: Mean Score in Mathematics by Country in PISA 2012 (OECD, 2013b)

3.3.2 Measure of Culture of Competitiveness

Culture is a highly subjective matter and hence hard to measure in numbers. In recent years international surveys have tried to make culture comparable across countries. Data is thus only available on a country level, which is generally justified by the fact that people from the same country share important determinants of the development of culture such as language and history. In this study a measure for a country's competitiveness is derived from a question from the WVS. The WVS is a global network of social scientists studying changing values and their impact on social and political life. The survey started in 1981 and consists of national surveys conducted in almost 100 countries using a common questionnaire. Random sampling is used in the countries to obtain representative national samples (Inglehart, 2014). From the WVS answers to the following statement are taken: "*Competition is good. It stimulates people to work hard and develop new ideas" vs. "Competition is harmful. It brings the worst in people*". Participants were asked to place their view about this statement on a scale from 1 to 10, where 1 means "*competition is good*" and 10 means "*competition is harmful*". 107,466 people were interviewed between 1989 and 2012. The data from all waves is aggregated on a country level and a simple average per country is calculated. The $Comp_c$ index is created by normalizing the data to take on numbers between 0 and 10, and reverse coded so that 10 is the most competitive country and 0 the least competitive. It is assumed that school children's competitiveness is captured by this aggregated measure since cultural values are shared by large groups or nations and are transmitted from parents to their children through generations. Looking at breakdowns of the data by age shows that the measure hardly changes if we only take the average of young participants or from old people. To confirm the time persistence of this cultural values, we calculate a $Comp_{ct}$ index per wave and run a panel data regression of the $Comp_{ct}$ index on country and time dummies. Performing F-tests on the time dummies proves that these are insignificant (p-value: 0.13). According to the created index $Comp_c$, competitive countries are those from Eastern Europe, the Balkan countries, the USA and Australia. Non-competitive countries are those from South-America and Western Europe. Asian countries are moderately competitive. For a full ranking of countries see Appendix 3.A.

To investigate further what factors determine the $Comp_c$ index, we calculate relevant correlations with country-level variables and discuss existing literature using the same measure. First, note that the correlation with the mean country score of PISA 2012 is negative and non-significant (−0.215), indicating that a competitive culture is not associated with a higher average achievement at school (see Figure 3.2). Hayward and Kemmelmeier

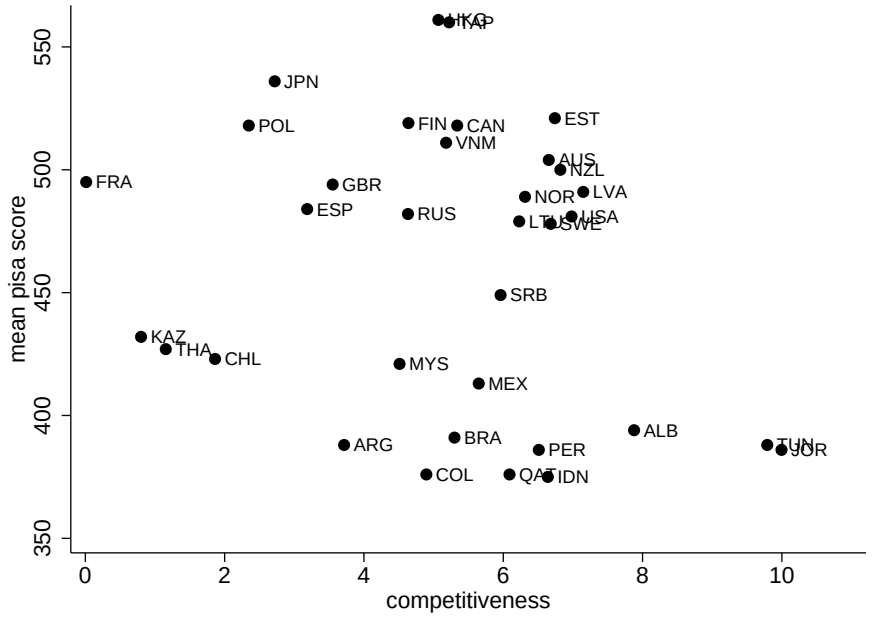


Figure 3.2: Relationship of Mean PISA Test Score and Competitiveness according to the WVS, Normalized to Values from 0 (Non-Competitive) to 10 (Competitive)

(2007) examine the structural and cultural roots of competitive attitudes using the $Comp_c$ measure including 81 countries. They find that the index is, if at all, negatively correlated to economic prosperity as measured by the per capita GDP or to economic freedom of a country as measured by the Heritage Foundation's Index of Economic freedom. The only significant finding of Hayward and Kemmelmeier (2007) is that competitive values are consistently correlated to Protestantism across societies as measured by the proportion of Protestants of the national population. According to the authors the Protestant culture is a value system that promotes the principles of free-market enterprise, and is hence likely to promote a competitive mindset. In our sample these are the Anglo-Saxon countries as well as Scandinavian countries. Hayward and Kemmelmeier (2007) find no correlation to individualism as opposed to collectivism measured by Hofstede (1984), who undertook an extensive survey about values at the workplace. In addition we calculate the correlation with the "Masculinity vs. Femininity" (MAS) measure developed by Hofstede (1984), which measures performance orientation as associated with masculine societies vs. cooperation orientation as associated with feminine societies. The correlation with this index is also small and insignificant (0.072). We argue that $Comp_c$ does thus not capture values measured by MAS such as performance orientation or free-market orientation and prosperity as measured by the GDP. Instead the WVS question used for $Comp_c$ does specifically mention the word "competition", so that it captures the aspect of social

comparison.[3]

Guiso, Sapienza, and Zingales (2003) study the impact of religion on economic attitudes, also using the $Comp_c$ measure under consideration in this paper. They find that Catholics and Protestants are in favor of competition, whereas Muslims and Hindus are strongly against it. Hindu countries (THA, VNM) and some Muslim countries (MYS, IDN) also score low in our sample. However, Jordan and Tunisia as non-Asian, but Muslim countries are obvious outliers, as well as France being Catholic but non-competitive.

3.3.3 Measure of Ability Grouping

The policy of AG implies that students are sorted into groups based on their ability or past achievement. These groups are then taught on different levels of difficulty. There are, however, several forms of AG that differ in their rigidity: First, there is the most rigid form: *countrywide ability tracking*. This means students are separated into different schools, usually based on achievement in primary school. Secondary schooling is then organized in two or three different tracks (schools). In this form of AG students are completely sealed off from students with different abilities. Second, there is *between-class grouping*, where students are separated into different classes within a school based on ability levels. And third, there is *within-class grouping*, where a class is divided into groups based on ability and achievement. This is commonly accomplished by assigning every member of the class to a particular group that they will be taught with during instruction in a particular subject. This is the least rigid form, since students still know and observe students with heterogeneous abilities within their class.

The strongest effect of AG is expected when students are grouped into different schools, so that reference point formation is only possible within the schools. Since this form of AG is implemented on a country level there is only little scope for regression analysis. Too few observations are available and a lot of other country-specific factors are likely to confound the analysis. Effects of AG are hardly ever found with this approach (Hanushek and Woessmann, 2006). Still, the same mechanisms should be at work when considering between-class grouping, the second most rigid form. If significant effects are found here, the effects in countries with rigid track formation should be even stronger. Considering between-class-grouping enables us to conduct the analysis on a school level, yielding many more observations and variation.

[3]Conducting the same analysis as done in this paper with MAS instead of COMP, yields insignificant results, indicating that the effect of AG does not depend on values measured by MAS. There is, however, a positive correlation of MAS with the average country score from PISA 2012, which illustrates the performance orientation measured by MAS.

"*Schools sometimes organize instruction differently for students with different abilities and interests in mathematics.* **Which of the following options describe what your school does for 15-year-old students?**" Please tick one box per row.

	For all classes	For some classes	Not for any classes
a) Mathematics classes study similar content, but at different levels of difficulty.	☐	☐	☐
b) Different classes study different content or sets of mathematics topics that have different levels of difficulty.	☐	☐	☐

Figure 3.3: Question on Ability Grouping in the PISA 2012 School Questionnaire

The PISA school principal questionnaire includes a question on ability grouping that is shown in Figure 3.3. From this question the variable AG_{sc} is constructed. This variable has the following six categories: (0) "not for any classes" for both a) and b); (1) "for some classes" for either a) or b) and "not for any classes" for the other; (2) "for some classes" for both a) or b); (3) "for all classes" for either a) or b) and "not for any classes" for the other; (4) "for all classes" for either a) or b) and "for some classes" for the other; (5) "for all classes" for both a) and b). Of all schools in the sample 16% have no AG (category 0), 13% are in category 1, 29% have some AG as in category 2, 16% of schools are categorized into 3, 14% in 4 and 11% of schools group all classes as in 5. Figure 3.4 shows the percentage of schools in the respective category of the variable AG_{sc} for all 34 countries included in the sample. The variable shows sufficient variation within and between countries. Remeber that only countries with a countrywide comprehensive school system are included in the sample as explained in Section 3.3.1. Countries with a relatively high percentage of grouped classes are (traditionally) English-speaking countries like Great Britain, the USA, New Zealand and Australia. Relatively little AG can, for example, be found in Scandinavian countries. The correlation between the amount of AG in a country (mean of AG_{sc} by country) and the culture of competitiveness is positive, but rather low and not significant (0.24).

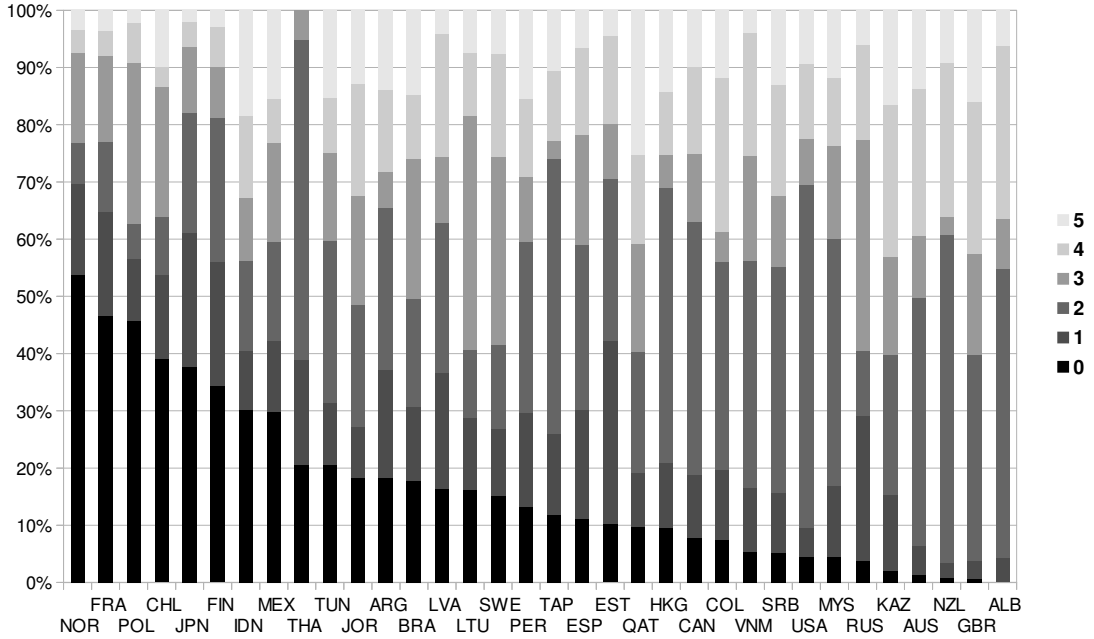


Figure 3.4: Share of Schools in a Country according to Categories of AG_{sc}

3.3.4 Control Variables

A standard set of control variables at the student and school level as found in many recent publications using PISA data is included (see e.g. Hanushek, Link, and Wößmann, 2013). In addition, some context related variables that might be correlated with our variable for AG are also added. Table 3.2 describes all control variables used in all following estimations.

Table 3.2: Description of Control Variables

Variable Name	**Description**
Student level:	
Age	Age of the student in years
Female	Dummy=1 if student is female
Grade Repetition	Dummy=1 if student ever repeated a grade
Grade	Grade of the student compared to modal grade for 15-year-olds in the country
Other Language at Home	Dummy=1 if student speaks a different language than the test language at home
Parents' Education	Highest completed level of education of both parents with categories: None (1), Primary School (2), Lower Secondary (3), Upper Secondary 1 (4), Upper Secondary 2 (5), University (6)
Books	Books at the home of the student (excluding school textbooks) with categories: 0-10 (1), 11-25 (2), 26-100 (3), 101-100 (4), 201-500 (5), more than 500 (6)

Table 3.2: (continued)

Variable Name	**Description**
Index of Socio-Economic Status (HISEI)	Index of the parents' socio-economic status, ranging from 0-100, taking into account their occupation and wealth
Class Size	Class size of the student's test language class
School level:	
Number of Students	Total student enrollment at the school
Private School	Dummy=1 if the school is a private school
Government Funding	Share of funding by the government
School Location	School location with categories: Village (1), Small Town (2), Town (3), City (4), Large city (5)
Math-Teacher Shortage	Dummy=1 if principal reports a shortage of math teachers
Student-Teacher-Ratio	Ratio of number of students to number of math-teachers at school
School Autonomy	Index on how much autonomy the school has regarding school budget, hiring and firing of teachers, teacher salary, courses offered etc.
Admission by Ability	Indicator on whether the school admits students based on academic record with categories: Never (1), Sometimes (2), Always (3)
Same Textbook	Dummy=1 if the school uses the same mathematics textbook for all classes

It is controlled for many factors that might determine whether a school practices AG or not, for instance the total number of student enrollment, school location, the type of school (private vs. public) and the share of government funding. Also racial and socioeconomic heterogeneity of a schools student body influence a schools decision to group (VanderHart, 2006). Including variables that control for this (e.g. *Other Language at Home*, *Books*, *Parents' Education*) ensures that there is no omitted variable bias, in the sense that the indicator for AG just picks up the effect of one of these variables. Controlling for whether the school admits students based on their prior achievement (*Admission by Ability*) is also important, since the student body at a school that undertakes this policy is a selection of high-ability students and AG in such a school probably has a lower effect.

The effect of AG on performance is not only driven by peer effects, but also by other factors like more appropriately paced instruction, smaller class size and focused curricula in ability segregated groups. For some of these factors it is controlled for by variables within the school characteristics vector, e.g. *Class Size*, which is usually smaller in schools where AG is used. Also, we control for *Same Textbook*, which indicates whether the school uses better suited curricula for different ability groups. Furthermore, we argue that if significant effects for an interaction of AG with competitiveness are found, there must be peer effects at work, since the variable $Comp_c$ is defined by social comparison. Table 3.3

shows correlations of AG_{sc} with important control variables. All the correlations are highly significant, but low in size.

Table 3.3: Pairwise Correlations of Ability Grouping and Selected Control Variables

	Ability Grouping		Ability Grouping
Class Size	-0.0536***	Number of Students	0.0173***
Books	-0.0370***	Private School	0.0139***
Index of Socio-Economic Status	-0.0065***	Admission by Ability	-0.0409***
Government Funding	0.0374***	Same Textbook	-0.0826***
School Location	-0.0135***		

Notes: Weighted by students sampling probability. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

More than fifty percent (58%) of the students have one missing value in at least one of the reported control variables. There is no pattern of missing values, but the values seem to be missing at random (MAR) in a non-monotone manner. Dropping all students with missing values would result in a substantial loss of observations and would lead to biased coefficients. As a solution missing data is imputed using the data of students with non-missing data as proposed by Woessmann (2003) and Ammermüller (2005). See Appendix 3.B for a detailed description of the imputation technique. Appendix 3.C provides summary statistics (mean and standard deviation) of student achievement and all imputed control variables.

3.4 Estimation Technique

The underlying model is an education production function framework, which typically explains student achievement by variables on the individual, the school and the country level (see e.g. Woessmann, 2003) resulting in a multi-level model. This model is augmented by the measure of competitiveness.

$$A_{isc} = \alpha + \beta_1\ AG_{sc} + \beta_2\ Comp_c + \beta_3\ AG_{sc} \times Comp_c + \mathbf{FB}_{isc}\gamma + \mathbf{S}_{sc}\delta + \mathbf{C}_c\kappa + \epsilon_{isc} \tag{3.1}$$

The dependent variable A_{isc} is math achievement of student i in school s and in coun-

try c as measured by PISA 2012.[4] The variable AG_{sc} is the indicator for AG as described in Section 3.3.3. The variable $Comp_c$ is the country level indicator for competitiveness as described in Section 3.3.2. To test whether the impact of AG varies with the competitiveness of students an interaction of AG_{sc} and $Comp_c$ is included. The vector $\mathbf{FB}_{isc}$ is a vector of the family background variables, $\mathbf{S}_{sc}$ a vector of the school characteristics and $\mathbf{C}_c$ is a vector of country characteristics. The error term is composed of errors at the individual student level, at the school level and at the country level:

$$\epsilon_{isc} = \eta_c + \eta_{sc} + \eta_{isc} \tag{3.2}$$

Here the country-specific error term η_c includes a set of cultural and educational factors for country c that cannot be measured, η_{sc} is a school-specific and η_{isc} an individual-specific error term. Since the purpose here is to find effects at the school level, country fixed effects μ_c can easily be included to control for unobserved country-specific factors, i.e. get rid of η_c. This also eliminates the variable $Comp_c$ because of perfect multicollinearity with the fixed effects. However, $Comp_c$ can stay in the interaction, which varies on a school level.

$$A_{isc} = \alpha + \beta_1 AG_{sc} + \beta_3\ AG_{sc} \times Comp_c + \mathbf{FB}_{isc}\gamma + \mathbf{S}_{sc}\delta + \mu_c + \epsilon_{isc} \tag{3.3}$$

The error term is now only composed of errors at the individual and at the school level, η_{sc} and η_{isc}. It is not possible to include school fixed effects, since this would eliminate the AG_{sc} variable. However, a wide set of school level variables is included, assuming that there are no unobserved school-specific effects left that are correlated with AG_{sc}. Despite the country fixed effects we still need the assumption of no unobserved cross-country heterogeneity that is related to the effect of AG on achievement for the identification of Equation (3.3). The only two channels discussed in the literature that determine how achievement is influenced by AG are peer effects and more appropriately paced instruction (Hanushek and Woessmann, 2006). Since the latter channel is unlikely to vary with culture, we assume that the coefficient β_3 captures the influence of culturally varying peer effects.

Equation (3.3) is estimated using ordinary least squares (OLS). To take into account the clustered nature of the data, where students are nested within schools and the schools are nested within countries, cluster-robust standard errors are used at the highest, namely

[4] PISA does not offer a single variable for student achievement, but 5 plausible values. Plausible values are random values drawn from a mathematically computed distribution of students' ability based on their test results and provide better estimates at the population level. Instead of one, there are thus five regressions to be computed for five different dependent variables. Results for coefficients and standard errors are averages of the results from the five plausible value regressions.

the country level. The sampling design of PISA is not completely random, which is why weights are used for every student consisting of the school weight and within-school weight to account for different sampling probabilities. The complex survey design of PISA also makes it necessary to use replication methods for computing the sample variance. PISA suggests Balanced Repeated Replication (BRR) with Fay's modification (OECD, 2005, pp.23), which is used here accordingly.[5]

The main interest is in the coefficients β_1 and β_3 of Equation (3.3). The coefficient β_1 can be interpreted as the change in average math achievement, if the variable AG_{sc} increases by 1 category for students in non-competitive countries (i.e. $Comp_c = 0$). The coefficient β_3 is the change in average achievement, if $Comp_c$ increases by one for students that are subject to AG as in category 1.

The regression might still suffer from selection bias, since good students could be attracted by schools that have a system with AG. This problem can be interpreted as an omitted variable bias, with innate ability being the omitted variable. This would result in η_{isc} being correlated with AG_{sc}. If this is the case, we would expect β_1 and β_3 to be positively biased. Some researchers (e.g. Ammermüller, 2005) argue that the problem of omitted ability does not matter in education production frameworks, since many proxy variables for ability are already included (e.g. parents' education, number of books at home, parents' occupation). The omitted variable issue shall still be considered in Section 3.6 as a robustness check.

3.5 Results

3.5.1 Pooled OLS wit Country Fixed Effects

The results from an OLS estimation of Equation (3.3), with and without the interaction of AG_{sc} with the $Comp_c$ index, are presented in Table 3.4. The estimated coefficients of all included control variables are given in Appendix 3.C. These coefficients are in line with previous research using PISA (e.g. Woessmann, 2003; Hanushek and Woessmann, 2014). Specification (1) and (3) in Table 3.4 include the variable AG_{sc} as the ordered categorical variable described in Section 3.3.3, and specification (2) and (4) in dummy coding with "no classes grouped" ($AG_{sc} = 0$) being the reference category. From the specifications with AG_{sc} in dummy coding we can conclude that these estimations are more meaningful, since the distances between the coefficients on the different categories of grouping are very different.

[5]For regression computing STATA is used with the PV module designed by Macdonald (2014).

Table 3.4: The Effect of Ability Grouping on Achievement (Pooled OLS)

Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.220	-3.457**		
	(0.596)	(1.441)		
Ability Grouping × Comp		0.615**		
		(0.292)		
AG = 1 (Some Classes Grouped)			-1.660	-10.500*
			(3.355)	(6.109)
AG = 2			-0.942	-21.558***
			(2.620)	(5.753)
AG = 3			0.188	-8.431
			(2.698)	(5.954)
AG = 4			-1.528	-16.579**
			(3.265)	(7.733)
AG = 5 (All Classes Grouped)			-1.878	-18.751**
			(3.709)	(8.689)
AG = 1 × Comp				1.911
				(1.468)
AG = 2 × Comp				4.226***
				(1.206)
AG = 3 × Comp				1.894
				(1.246)
AG = 4 × Comp				3.103*
				(1.662)
AG = 5 × Comp				3.385**
				(1.689)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,558	10,558	10,558	10,558
Student obs.	249,968	249,968	249,968	249,968
avrg. R^2	0.49	0.49	0.49	0.49

Notes: Dependent variable: PISA 2012 math test score. Reference Category is 'no grouping in any classes'. Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: age, female, parents' education, hisei, grade, grade repetition, books at home, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, math-teacher shortage, same textbook, admission by ability, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

To interpret coefficients note that the PISA math score was normalized to have a mean of 500 and a standard deviation of 100 across OECD countries in 2003. The first and third specification in Table 3.4 show no significant effect of AG on achievement when the interaction term with culture is not included. This corresponds to previous research that did not find effects of AG on average performance. However, the specifications that include the WVS measure for competitiveness show that culture *does* matter. Here AG has a significant negative effect in countries with low competitiveness, but a positive effect in competitive countries. For example, from specification (4) we see that in a country scoring 0 on the $Comp_c$ index, AG in all classes (as in category 5) compared to AG in no classes reduces achievement on average by 19 score points. In a country scoring 10

on the $Comp_c$ index AG in all math classes increases average achievement by about 15 score points. Finally, for a medium competitive country of $Comp_c$ = 5 there is no effect of grouping in all classes. Specification (4) also shows that already grouping in "some classes" as in category 2 of the variable AG_{sc} has a strong effect. It leads to a decrease of 22 points of average student achievement in non-competitive countries ($Comp_c$ = 0) and an increase of 21 points in competitive countries ($Comp_c$ = 10). Schools reporting that *some* classes are grouped might, for instance, be comprehensive schools that have remedial classes for particularly bad students or extra math classes for particularly good students. The estimated coefficients are relatively large compared to estimated effects of school inputs in the PISA literature. For example Fuchs and Wößmann (2008) find that 1000 hours of extra instruction time per year lead to an increase in average achievement by 5 score points and that students at a publicly managed school perform on average 19 score points lower than students at privately managed schools.

3.5.2 Quantile Regression

To test whether low or high-ability students suffer or gain more from AG, quantile regressions with country fixed effects according to Koenker (2004) are run. This enables us to see whether there is heterogeneity in the effects of grouping across the conditional achievement distribution. Since it is controlled for all kinds of family and student characteristics, the conditional achievement distribution should be strongly correlated with innate ability, or more precisely the part of ability that is not correlated to the measured student characteristics (for a similar approach see Woessmann, 2008). We will thus from now on refer to the conditional achievement distribution as the ability distribution. Since it can be assumed that the distribution of innate ability is constant across countries, we do not have to worry about the different achievement distributions in different countries. Quantile regressions estimate the effect of grouping on student achievement for students at different points on the ability distribution. Table 3.5 reports the coefficients on AG_{sc} and $AG_{sc} \times Comp_c$ in dummy coding for the quantiles ranging from 0.1 to 0.9. Parente and Santos Silva (2013) show that the quantile regression estimators are also consistent when the error terms are correlated within clusters.

For students at all quantiles we find significant negative effects in non-competitive countries and significant positive effects in competitive cultures. Focusing on the coefficients on the AG = 5 dummy, which indicates the change in average achievement if all classes in the school are grouped compared to no grouped classes, we see that the effect of AG is biggest for students at the median and becomes smaller the further away we go in

Table 3.5: Quantile Regressions

Variables	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AG=1	-4.474**	-6.798***	-9.565***	-11.143***	-12.873***	-9.226***	-11.423***	-6.747***	-5.529**
	(2.199)	(1.800)	(1.705)	(1.685)	(1.692)	(1.716)	(1.813)	(1.833)	(2.288)
AG=2	-11.793***	-16.344***	-16.695***	-19.346***	-21.184***	-21.969***	-23.391***	-24.219***	-24.319***
	(1.974)	(1.616)	(1.530)	(1.512)	(1.518)	(1.540)	(1.627)	(1.645)	(2.054)
AG=3	-4.880**	-6.572***	-6.752***	-7.103***	-10.642***	-8.205***	-8.212***	-6.324***	-4.553*
	(2.266)	(1.855)	(1.757)	(1.736)	(1.743)	(1.768)	(1.868)	(1.889)	(2.358)
AG=4	-7.331**	-11.384***	-17.530***	-18.095***	-22.570***	-19.798***	-19.887***	-17.919***	-16.118***
	(2.858)	(2.339)	(2.215)	(2.190)	(2.198)	(2.229)	(2.356)	(2.382)	(2.973)
AG=5	-13.210***	-15.748***	-16.925***	-18.020***	-21.482***	-14.768***	-13.483***	-10.524***	-9.262***
	(3.199)	(2.619)	(2.480)	(2.451)	(2.461)	(2.496)	(2.638)	(2.667)	(3.329)
AG=1 × Comp	0.211	1.099***	1.756***	1.984***	2.519***	1.664***	1.912***	0.695*	0.873*
	(0.439)	(0.360)	(0.341)	(0.337)	(0.338)	(0.343)	(0.362)	(0.366)	(0.457)
AG=2 × Comp	2.097***	3.347***	3.595***	3.975***	4.347***	4.392***	4.479***	4.435***	4.760***
	(0.372)	(0.304)	(0.288)	(0.285)	(0.286)	(0.290)	(0.306)	(0.310)	(0.387)
AG=3 × Comp	1.377***	1.823***	1.812***	1.705***	2.360***	1.854***	1.757***	1.097***	0.689
	(0.434)	(0.356)	(0.337)	(0.333)	(0.334)	(0.339)	(0.358)	(0.362)	(0.452)
AG=4 × Comp	1.304**	2.231***	3.335***	3.300***	4.345***	3.941***	3.741***	3.091***	3.219***
	(0.515)	(0.421)	(0.399)	(0.394)	(0.396)	(0.402)	(0.424)	(0.429)	(0.536)
AG=5 × Comp	1.922***	2.665***	3.180***	3.357***	3.841***	2.680***	2.437***	1.706***	1.982***
	(0.563)	(0.461)	(0.436)	(0.431)	(0.433)	(0.439)	(0.464)	(0.469)	(0.586)
Student controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34	34	34	34	34	34
School obs.	10,558	10,558	10,558	10,558	10,558	10,558	10,558	10,558	10,558
Student obs.	249,968	249,968	249,968	249,968	249,968	249,968	249,968	249,968	249,968
Avrg. pseudo R^2	0.22	0.25	0.27	0.28	0.30	0.31	0.32	0.32	0.33

Notes: Dependent variable: PISA 2012 math test score. Quantile regression weighted by students sampling probability. Standard errors given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, books at home, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

each direction. In very non-competitive cultures ($Comp_c = 0$) AG in all classes decreases achievement by 13 score-points for students with very low ability at the 0.1 quantile and decreases achievement by 9 score-points for high-ability students at the 0.9 quantile. In very competitive cultures ($Comp_c = 10$) AG in all classes increases achievement by 19 score-points for low-ability students at the 0.1 quantile and by roughly the same amount for high-ability students at the 0.9 quantile. The median regression can be viewed as a test of the OLS regression that is robust against outliers (Woessmann, 2008). Here it clearly supports the results of the least-squares regression with coefficients on $AG = 5$ and $AG = 5 \times Comp$ (all classes grouped) being a bit bigger.

Figure 3.5 shows the impact of $AG = 5$ (all classes grouped compared to no classes grouped) on achievement in score-points for non-competitive countries ($Comp_c = 0$, in light grey) and competitive countries ($Comp_c = 10$, in dark grey) across the quantiles. It shows both the estimates including all the control variables from previous estimations

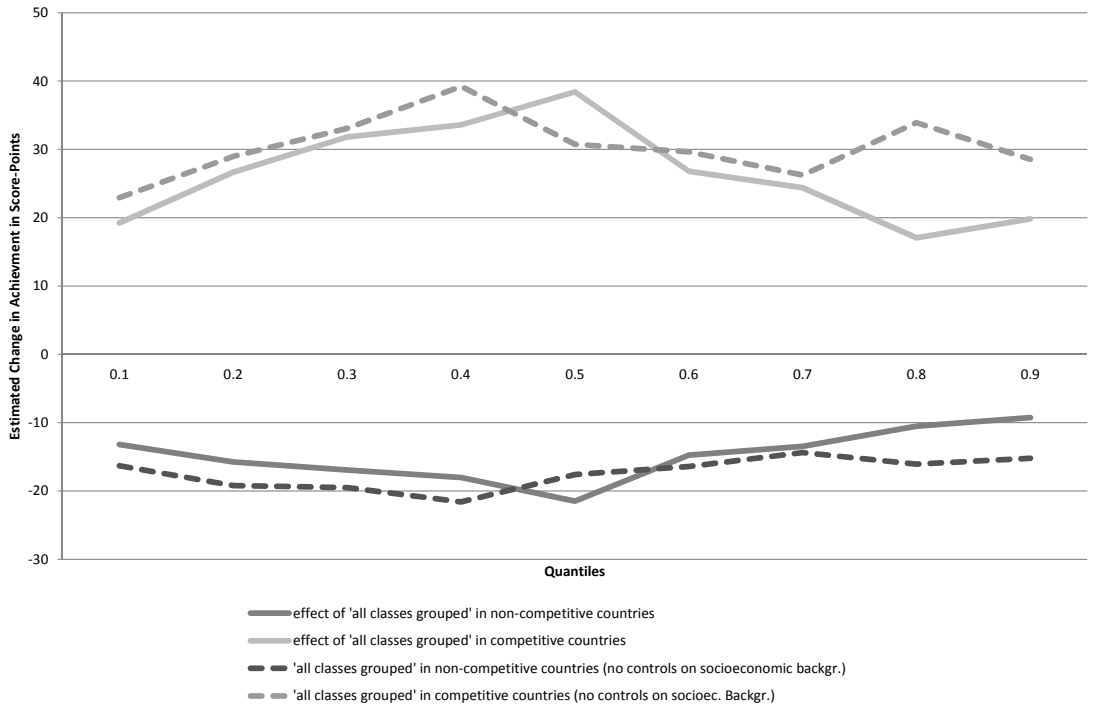


Figure 3.5: Estimated Effect of "All Classes Grouped" at Different Quantiles of the Conditional Achievement Distribution for Competitive and Non-Competitive Cultures

(see Table 3.5) as well as from quantile regressions without the control variables that might be proxies for ability[6] (dashed lines). This is done as a robustness check to ensure that the correlation of the conditional achievement distribution with innate ability is not eliminated by these control variables. The conditional achievement distribution from estimations excluding these control variables should then be highly correlated to innate and nurtured ability.

Figure 3.5 shows that the influence of AG is generally smaller at the tails of the ability distribution. A possible explanation for this result is that medium-ability students are confronted with the biggest change in their social position when they are sorted into ability based groups. While they are mediocre under comprehensive schooling they are either among the best or worst students in a two-track system. Without controlling for variables on socioeconomic background the effects are generally the same, but slightly bigger. This might be because also nurtured ability is important for the effect of AG on achievement. Another explanation is that these socioeconomic variables are correlated with AG_{sc}, so that the coefficient on AG_{sc} in the regressions without these controls catches some of their effects.

[6]i.e. without *books at home*, *hisei* and *parents' education*

The results from the quantile regressions also yield insights on the effect of grouping on the variance or inequality of achievement. Since only little variation is found across the quantiles, variance effects of AG should be small. In non-competitive countries the variance in grouped schools is larger than in comprehensive schools, since low-ability students lose more from grouping than high-ability students. In competitive countries the variance is also slightly bigger under grouping, since high-ability students gain more than low-ability students. Also a regression using the standard deviation of achievement per school as dependent variable supports the result that between-class grouping has little influence on the inequality of student achievement (see Appendix 3.E).

3.6 Instrumental Variables

There is a possibility that the variable AG_{sc} is endogenous. This is because school choice of students (or their parents) might be affected by whether a school does or does not group by ability. For example, good students might be attracted by schools that have groups for high-ability students, since this gives them the opportunity to study at a higher difficulty without being slowed down by low-ability students. In this case the above estimates for AG_{sc} would be biased upwards. This problem can also be interpreted as an omitted variable bias, as in Betts and Shkolnik (2000), with innate ability being the omitted variable. Since innate ability is probably positively correlated with AG_{sc}, an endogeneity problem arises. In order to address this problem an instrumental variable approach is suggested using as an instrument a variable that yields information on whether students have a choice between different schools.

The instrument suggested is data from a question from the PISA 2012 school questionnaire about how many schools the school is competing with in the region. From this question a variable $Schoolcomp_{sc}$ is constructed that takes on the value 0 if the school is not competing with any other school, 1 if the school is competing with one other school and 2 if there are two or more schools the school competes with. Figure 3.6 illustrates that there is a lot of variation in this variable between and within countries. Naturally more availability of schooling is found in countries that are more densely populated.

A positive correlation between $Schoolcomp_{sc}$ and AG_{sc} is expected for two reasons. First, the availability of schooling in the region is a natural predictor of self-selection since no selection can take place when students do not have a choice between schools. Therefore the effect of AG on students that do not have a choice between schools can be compared with those that have a choice. Second, school competition might also affect a schools decision to group classes or not to group. If a school is competing for students with other

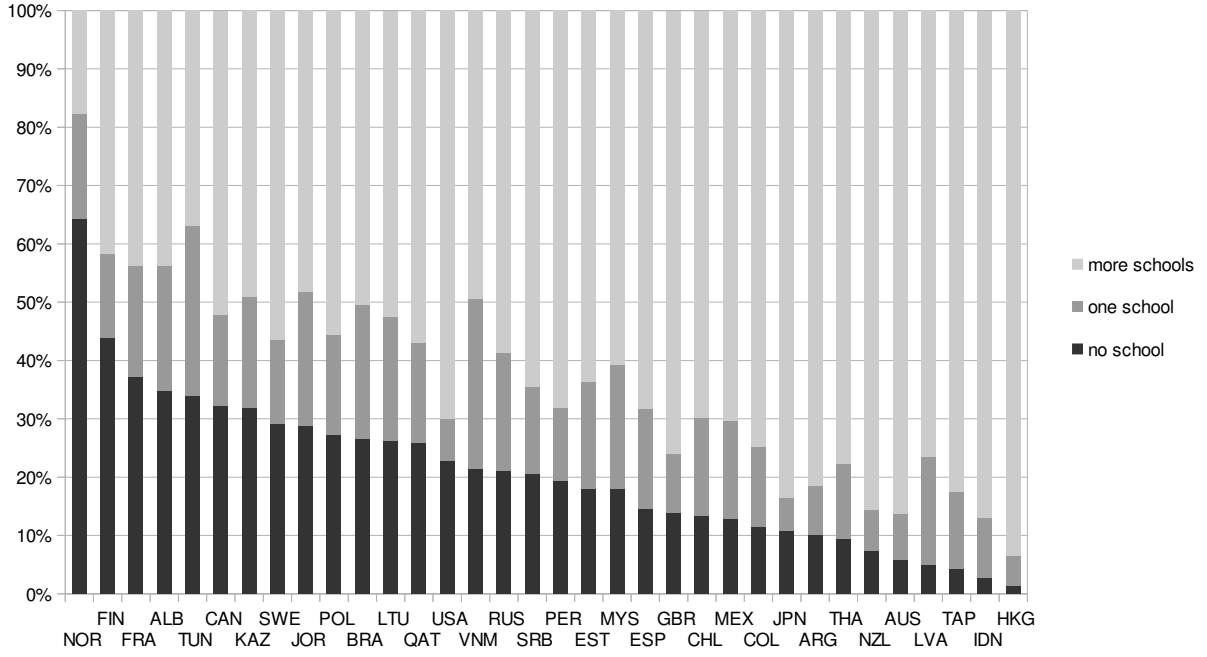


Figure 3.6: Number of Schools a School Competes with in PISA 2012

schools, it might rather offer ability-grouped classes in order to attract high-ability students.

Furthermore, we argue that the instrument is valid in terms of it not being correlated with the dependent variable *achievement*. First of all, the fact that there are more or less schools in a region is mostly exogenously given from historic and geographical reasons. One could argue that more schools open in regions where residents' education is high, and students are expected to be of high ability. It can be shown, however, that the correlation between $Schoolcomp_{sc}$ and $Books_{isc}$ (a proxy for students ability and family background) is very low (0.05). In fact *location* has the highest correlation with $Schoolcomp_{sc}$ (0.39), showing that the bigger the town, the more schools compete with each other. This underlines the exogenous character of $Schoolcomp_{sc}$. See also Currie and Moretti (2003) for arguments in favor of exogeneity of the number of schools in a given area. Furthermore, it might be argued that school competition improves a schools' quality, leading to a positive correlation between achievement and $Schoolcomp_{sc}$. However, research shows that there is no significant positive link between active school choice and achievement. These studies use randomized lotteries due to the highly selective nature of students who chose their school (see Musset, 2012, p.25).

In the IV approach the variable AG_{sc} is not used in dummy coding, since more instruments would then be needed.[7] Still, the endogenous variable AG_{sc} appears twice in

[7]Using an approach with AG_{sc} in dummy coding and only instrumenting the dummy for $AG = 5$ (all classes grouped) yields roughly the same results as those presented here.

the main regression. Once on its own and once in the interaction with $Comp_c$. Therefore there are two endogenous variables in the regression for which we need two instruments. According to Wooldridge (2002, pp.121) the natural instrument for an interaction is to substitute the endogenous variable in the interaction with the instrument. Thus, $Schoolcomp_{sc} \times Comp_c$ is the instrument used for $AG_{sc} \times Comp_c$.

Table 3.6 yields the results of the first-stage regressions from a two-stage-least squares approach as well as for a baseline model that does not include the $Comp_c$ interaction. The results illustrate that $Schoolcomp_{sc}$ is positively correlated with the endogenous variable AG_{sc} in the baseline regression. Once the interaction with $Comp_c$ is included the coefficient on $Schoolcomp_{sc} \times Comp_c$ is positively significant, suggesting that the more competitive a country, the more do students choose ability-grouped schools. This indicates that the more competitive a student's attitude, the more do they actively seek a competitive environment, i.e. ability-grouped schools, if they have the choice. Students in non-competitive countries do not seem to actively choose comprehensive or ability-grouped schools.[8]

Table 3.6: First-Stage Regressions

Variables Dep. Variable	(1) Ability Grouping	(2) Ability Grouping	(3) Ability Grouping × Comp
Schoolcomp	0.106***	-0.139*	-0.653*
	(0.041)	(0.077)	(0.389)
Schoolcomp × Comp		0.049***	0.283***
		(0.018)	(0.109)
Student controls	Yes	Yes	Yes
School controls	Yes	Yes	Yes
Country FE	Yes	Yes	Yes
Country obs.	34	34	34
School obs.	10,534	10,534	10,534
Student obs.	249,505	249,505	249,505
R^2	0.09	0.10	0.29
Robust Fstat	6.74***	9.52***	9.44***
Hausman	0.42	3.88	

Notes: Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, books at home, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

[8]Appendix 3.D also lists results for the cluster-robust OLS regression on male and female subgroups. Conducting the IV analysis only on male students, shows that F-tests on the first stage are higher, indicating more selective behavior of male students. The OLS analysis on these subgroups, however, shows that male and female students are equally affected by AG.

The cluster-robust F-statistic on the excluded instruments in the baseline specification (1) is too low for an IV approach. Also in the model including the interaction with $Comp_c$ (specification (2) and (3)) the F-statistic is just below 10, the level which is usually recommended as proof of strong instruments. In addition, a Hausman test is conducted, which is a test of the exogeneity of AG_{sc}. The chi-squared statistic for the significant error terms of the first-stage regressions included in the OLS regression is not significant in either model. This indicates that there is no evidence of endogeneity of AG_{sc}. However, the Hausman test is only as good as the instrument used and in case of a weak instrument might fail to diagnose endogeneity correctly (Hahn and Hausman, 2003). To increase the power of the instrument first-stage regressions on different subgroups of the population are conducted (similarly in Figlio and Page, 2002). This might increase the power since the monotonicity of the instrument might not be given. Probably not all types of students, high or low-ability, have equal selection behavior. Theoretical predictions from Thiemann (2017) suggest that high-ability students profit from ability-grouped schools, while low-ability students profit from comprehensive schooling. Likewise different selection behavior is expected from different ability groups. As a proxy for student's ability the variable $Books_{isc}$ is used, which indicates in six categories how many books there are at the home of a student.[9] This variable serves as an indicator of parents' education and socio-economic status and should be highly correlated with student's ability, since ability depends to a high degree on genes as passed on by parents and nurture at home (Plomin et al., 1997).

Table 3.7 shows the first-stage results with the $Comp_c$ interaction for students from the six different categories of the variable $Books_{isc}$. Only results for the regression with AG_{sc} as dependent variable are shown (results for the regression with $AG_{sc} \times Comp_c$ are similar). It can be seen that selection only takes place among students with high or medium ability. Again, we observe that students from competitive countries select into ability-grouped schools. For high-ability students from non-competitive cultures we now also find evidence of selection behavior into comprehensive schools. The coefficient on the interaction $Schoolcomp_{sc} \times Comp_c$ is significant even at the 1% level, suggesting that selection behavior in competitive countries is stronger. The lack of significance for students with low ability can be explained with selection criteria of schools. Bad students might thus not even have a choice between schools, since they are not admitted to certain private schools. Also, parents of students from the two lowest subgroups might lack knowledge about strategic school choice or they lack ambition with respect to their child's education. In addition, since $Books_{isc}$ is a variable that also captures the socio-

[9]Students answered the question on how many books there are at their home themselves. To illustrate the numbers of the six categories pictures of bookshelves were shown. It was also mentioned that schoolbooks should not be included (OECD, 2012).

Table 3.7: First-Stage Regressions on Sub-Samples

Variables Sample	(1) Books>500	(2) Books201-500	(3) Books101-200	(4) Books26-100	(5) Books11-25	(6) Books0-10
Schoolcomp	-0.300**	-0.243**	-0.165*	-0.168*	-0.049	-0.070
	(0.123)	(0.116)	(0.094)	(0.086)	(0.075)	(0.105)
Schoolcomp × Comp	0.093***	0.077***	0.072***	0.065***	0.019	0.030
	(0.026)	(0.024)	(0.020)	(0.019)	(0.018)	(0.024)
Student contr.	Yes	Yes	Yes	Yes	Yes	Yes
School contr.	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34	34	34
School obs.	5,531	6,909	8,476	9,944	9,564	9,305
Student obs.	13,163	23,374	32,956	67,487	52,877	59,648
Avrg. R^2	0.18	0.17	0.17	0.12	0.08	0.06
Robust Fstat	14.72***	13.57***	26.22***	16.40***	1.25	2.80
Hausman	0.80	0.66	1.06	3.11	5.94*	2.17

Notes: Dependent variable: Ability Grouping. Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

economic status of parents, they might not have the money to send their students to expensive private schools. The F-statistic for the significance of the instruments is well above ten in the first four subgroups of Table 3.7, which is a good foundation for an IV regression on these subsamples. Performing Hausman tests of endogeneity for the different subgroups, indicates that we can reject endogeneity of the variable AG_{sc} (see Table 3.7). None of the Hausman tests in the first four subgroups is significant, suggesting that the OLS estimates are the true estimates. In these four subgroups selection does seem to take place, but either to such a little extent that the OLS estimates are not biased or the inclusion of control variables on student background renders the omitted variable problem non-existent. As for the lowest two subgroups (books 11-25, books 0-10), the first-stage regression suggests that there is no self-selection of students. Therefore, OLS estimates yield the true estimates also for these subgroups. The results from the OLS and the quantile regression in Section 3.5 can thus be considered as robust to endogeneity and unbiased.

As a robustness check we repeat the IV analysis of Equation (3.3) without including student level controls on family background[10]. This is done to verify that our variable for grouping would be endogeneous in a regression without any proxy for innate ability. The Hausman test now yields significant results (chi-squared statistic: 8.59**; p-value:

[10]i.e. without *books at home*, *hisei* and *parents' education*

0.0136) indicating that we can reject exogeneity of AG_{sc}. This shows that the inclusion of family background variables, as done in all our regressions, can proxy effectively for unobserved innate ability such that the Hausman test is insignificant.

3.7 Further Robustness Checks

Several additional robustness checks are conducted. First, a cluster-robust OLS regression as in Equation (3.3) is conducted for the dependent variable *science* and *reading* achievement (see Appendix 3.F). The regressions yield roughly the same results as the reported ones in Section 3.5.1 with math achievement. For science significance is even stronger, for reading less strong. In addition, the OLS analysis is run with different definitions of the AG_{sc} variable. For instance, a question from the school questionnaire on *within-class* grouping can be included to define a variable $AG2_{sc}$ that takes on the following values: 0 if "not for any classes" was ticked both for between-class grouping and for within-class grouping; 1 if "not for any classes" was ticked for between-class grouping; 2 if between-class grouping is operated for "some classes" and 3 if between-class grouping is operated in "all classes". The results for the cluster-robust OLS analysis are given in Appendix 3.F. Again, results are similar to those reported in Section 3.5.1. Direction and significance of the effects are the same, only the coefficients are a bit smaller in size once $AG2_{sc}$ is considered. Within-class grouping has no significant effects.

For the results presented in the main part of this paper we decided to drop all observation of first and second generation immigrants, since national culture of the test country cannot be assigned to them. Since the effect of competitiveness is assumed to work via peer effects, the immigrant population could still matter in the sense that native students are affected by the performance of the immigrant students in their class. To account for this we repeat the regression as in Equation (3.3) including the immigrant population, but controlling for their immigrant status, also in an interaction with AG_{sc}. The results as given in Appendix 3.F.3 show that the coefficients on AG_{sc} and $AG_{sc} \times Comp_c$ do not change compared to those reported in Section 3.5.1. The coefficients on the controls for immigrants are insignificant.

An estimation technique very often used in educational research is multi-level-analysis, i.e. a random effects model that takes into account the different levels of observation of the data, namely student, school and country level. Since the interest of this paper is only in the effects at the school level, so far only the OLS analysis was presented with country fixed effects and standard errors adjusted for the country clusters. However, the results of a random effects model shall be given as a robustness check (see Appendix 3.F).

The results are qualitatively the same as the OLS results presented in Table 3.4, with coefficients on AG_{sc} and $AG_{sc} \times Comp_c$ being a bit bigger in size. The regression results also show that the model can explain almost 70% of the between-school variation, but only 12% of the within-school variation.

3.8 Conclusion

The analysis of school level PISA 2012 data has shown that culture, or more precisely competitiveness, *does* matter for the effect of AG on student performance. Particularly, we find evidence for AG being detrimental in non-competitive cultures, but beneficial in competitive cultures. Students at the tails of the ability distribution are generally less affected than those closer to the median. The effect of AG on the variance of achievement is not significantly different from zero.

The positive effect of AG in competitive cultures supports the idea that being surrounded by students of similar ability can be more motivating than being in a class with students of heterogeneous abilities. This positive effect of AG can be explained by the model from Thiemann (2017) including a non-linear value function, thereby modeling diminishing sensitivity to the reference point. For instance the value function of Tversky and Kahneman (1979) is convex below the reference point, indicating that being just below the reference point induces a higher motivation than being further away. Another explanation can be the existence of a participation constraint (Thiemann, 2017). Students that give up because of being too far away from the reference point are mainly a problem in comprehensive schools, where abilities are very heterogeneous. Under AG, however, the reference point is usually close enough to drive students to perform. Furthermore, competitive students in ability-grouped schools might be incentivized by the chance of being promoted to a higher track, if they perform among the best of the group. This possibility has not been considered in the theoretical model and is subject to further research.

In non-competitive cultures evidence is found for students losing under AG, especially medium to low-ability students. The latter coincides with theory and can be explained by students in lower tracks having a lower reference point than under comprehensive schooling. Especially the relatively good students in the low track are not motivated anymore, since they have no-one to look up to. The overall detrimental effects of AG in non-competitive cultures could also be due to some kind of "competition-aversion", which is not covered by the theory of Thiemann (2017). The model includes the possibility that students are non-competitive in the sense that they do not get any utility or disutility from social comparison. Then students' utility increases only in own performance. It might

be possible, however, that relative performance feedback has a discouraging effect. For example students that feel comfortable being mediocre in a comprehensive school, would find themselves being a bad student in a high track of a grouped system. While this might drive competitive students to perform higher it might demotivate non-competitive students, because of too high expectations and pressure to perform. Correspondingly the IV approach has shown that competitive students actively seek more competitive environments (ability-grouped schools), whereas students from non-competitive cultures avoid this.

All in all the analysis has provided an important contribution to the existing literature by showing that there is a significant effect of AG on student performance once we distinguish between competitive and non-competitive cultures. This reveals that school systems have to be designed taking into account the culture in a given country. However, with field data from PISA it is hard to investigate the structure of incentives that drives students to perform at a certain level. For further research a laboratory experiment might be useful to disentangle the channels that drive subjects to perform and test the theoretical hypotheses in an environment that closely matches the model from Thiemann (2017). In an experiment confounding factors can be eliminated and factors considered in theoretical models (loss aversion, individual reference points, competitiveness) can be tested for directly.

Appendix

3.A Measure of Competitiveness from WVS

Country	Code	Competitiveness
Jordan	JOR	9.995
Tunisia	TUN	9.79
Albania	ALB	7.878
Latvia	LVA	7.148
USA	USA	6.980
New Zealand	NZL	6.819
Estonia	EST	6.740
Sweden	SWE	6.681
Australia	AUS	6.653
Indonesia	IDN	6.639
Peru	PER	6.510
Norway	NOR	6.312
Lithuania	LTU	6.228
Qatar	QAT	6.09
Serbia	SRB	5.961
Mexico	MEX	5.648
Canada	CAN	5.338
Brazil	BRA	5.299
Taiwan	TAP	5.222
Vietnam	VNM	5.179
Hong Kong	HKG	5.068
Colombia	COL	4.895
Finland	FIN	4.638
Russia	RUS	4.631
Malaysia	MYS	4.511
Argentina	ARG	3.714
United Kingdom	GBR	3.550
Spain	ESP	3.184
Japan	JPN	2.718
Poland	POL	2.346
Chile	CHL	1.862
Thailand	THA	1.155
Kazakhstan	KAZ	0.8
France	FRA	0.012

Notes: Reverse coded and normalized to values from 0 (non-competitive) to 10 (competitive).

3.B Missing Values

Including all control variables would result in a loss of almost 60% of the data if observations with missing values are dropped. 40% of the observations have one missing value, more than 19% of the observations even more. There is no pattern of missing values, but the values seem to be missing at random (MAR) in a non-monotone manner. Most values are missing for the variable *classsize*.

Table 3.8: Summary Statistics

Variable	Mean	Std. Dev.	N
Achievement	447.049	101.297	252,921
Age	15.805	0.292	252,808
Female	0.509	0.5	252,921
Grade Repetition	0.17	0.376	235,239
Other Language at Home	0.135	0.342	248,020
Parents' Education	4.017	1.803	249,324
HISEI	44.479	23.091	235,663
Books at Home	2.645	1.37	248,971
Grade	-0.273	0.727	252,684
Class Size	29.796	9.945	153,976
Number of Students	970.656	795.997	236,018
Private School	0.203	0.402	250,628
Math-Teacher Shortage	0.179	0.383	248,438
Student-Teacher-Ratio	153.548	123.374	224,337
School Location	2.986	1.249	250,378
Government Funding	78.099	33.317	228,977
School Autonomy	-0.048	1.114	250,882
Admission by Ability	2.033	0.894	248,645
Same Textbook	0.737	0.44	245,911

Dropping all students with missing values would result in substantial loss of observations and would lead to biased coefficients. As a solution we impute missing data using the data of students with non-missing data as proposed by Woessmann (2003) and Ammermüller (2005). Unlike using country-by-wave-means for the missing values this does not "distort covariances and intercorrelations between variables" (Schafer and Graham, 2002, p. 159) or introduce bias and understate variability (Horton and Kleinman, 2007, p. 80). Following Woessmann (2003, p.169) the technique works as follows: "For each student i with missing data on a specific variable M, a set of 'fundamental' explanatory variables F with data available for all students is used to impute the missing data. Let S denote the set of students j with available data for M. Using the students in S, the variable M was regressed on F:

$$M_{j \in S} = F_{j \in S}\phi + \epsilon_{j \in S} \tag{3.4}$$

For M being a discrete variable, OLS estimation was used for the regression. For M being a dichotomous (binary) variable, a probit model was used. If M was originally (before deriving dummies) a polychotomous qualitative variable with multiple categories, an ordered-probit model was estimated. The coefficients ϕ from these regressions and the data on F_i were then used to impute the value of M_i for the students with missing data:

$$\widetilde{M}_{j \notin S} = F_{j \in S}\phi \tag{3.5}$$

For the probit models, the estimated coefficients were used to forecast the probability of occurrence associated with each category for the students with missing data, and the category with the highest probability was imputed."

As fundamental variables that are complete for almost the whole data set we use student's *age*, *female*, *parents' education*, *wealth*, the *school location*, *GDP per capita* (World Bank, 2014b) and *public spending on education* (World Bank, 2014a). With these fundamental variables values for *grade repetition, other language at home, hisei, books at home, number of students, private school, math-teacher shortage, government funding, student-teacher-ratio, class size, school autonomy, admission by ability, same textbook* and *grade* are imputed. The small amount of missing data within F was imputed by taking the average value at the school level.

3.C Summary Statistics and Coefficients of Control Variables

Variables	Mean	Std. Dev.	Min.	Max.	Student Obs.
Math achievement	446.698	101.206	19.793	924.84	251,972
Ability Grouping	2.285	1.561	0	5	251,972
Student characteristics:					
Age	15.805	0.292	15.17	16.33	251,919
Female	0.508	0.5	0	1	251,972
Index of Socio-Economic Status (HISEI)	44.173	22.656	-0.652	88.960	250,774
Grade Repetition	0.163	0.35	0	1	251,914
Other Language at Home	0.14	0.343	0	1	251,681
Class Size	29.848	8.137	0	200	251,972
Grade	-0.254	0.700	-3	3	251,972
Parents' education:					
None	0.028	0.165	0	1	251,969
Primary School	0.1	0.3	0	1	251,960
Lower Secondary	0.124	0.33	0	1	251,969
Upper Secondary 1	0.04	0.195	0	1	251,969
Upper Secondary 2	0.411	0.492	0	1	251,897
University	0.296	0.457	0	1	251,969
Books at home:					
Books 0-10	0.259	0.438	0	1	251,835
Books 11-25	0.245	0.43	0	1	251,835
Books 26-100	0.272	0.445	0	1	251,835
Books 101-200	0.111	0.314	0	1	251,835
Books 201-500	0.074	0.262	0	1	251,835
Books > 500	0.039	0.194	0	1	251,972
School characteristics:					
Number of Students	973.801	792.23	1	11483	251,972
Private School	0.203	0.402	0	1	251,972
Math-Teacher Shortage	0.178	0.381	0	1	251,972
Student-Teacher-Ratio	154.462	118.86	0.5	2,311	251,585
Government Funding	78.237	32.171	0	116.302	251,810
School Autonomy	-0.037	1.11	-2.872	1.604	251,972
Admission by Ability	2.028	0.898	0.148	3	251,914
Same Textbook	0.737	0.438	0	1	251,971
School location:					
Village (< 3,000)	0.152	0.359	0	1	251,972
Small Town (3,000-15,000)	0.202	0.401	0	1	251,972
Large Town (15,000-100,000)	0.266	0.442	0	1	251,972
City (100,000-1,000,000)	0.258	0.438	0	1	251,972
Large City (>1,000,000)	0.121	0.326	0	1	251,972

	(1)		(2)	
Variables	Coefficients	Std.Error	Coefficients	Std.Error
Ability Grouping	-0.220	0.596	-3.457**	1.441
Ability Grouping × Comp			0.615**	0.292
Student characteristics:				
Age	0.478	1.234	0.465	1.234
Female	-13.900***	0.780	-13.887***	0.782
Index of Socio-Economic Status (HISEI)	0.597***	0.025	0.596***	0.025
Grade Repetition	-33.074***	1.517	-32.994***	1.506
Other language at home	1.276	2.314	1.237	2.298
Class size	0.561***	0.068	0.562***	0.068
Grade	20.715***	1.128	20.765***	1.124
Parents' education:				
Primary school	6.462***	2.015	6.464***	2.016
Lower secondary	5.402***	1.734	5.412***	1.737
Upper secondary 1	6.340**	2.491	6.410***	2.485
Upper secondary 2	6.912***	2.030	6.955***	2.027
University	19.342***	2.093	19.385***	2.090
Books at home:				
Books 11-25	7.696***	0.977	7.664***	0.976
Books 26-100	25.073***	1.133	25.049***	1.134
Books 101-200	36.714***	1.654	36.637***	1.644
Books 201-500	59.923***	1.832	59.930***	1.825
Books > 500	46.294***	2.670	46.285***	2.651
School characteristics:				
Number of student	0.007***	0.001	0.007***	0.001
Private school	-7.080**	3.038	-7.468**	3.012
Math-teacher shortage	-8.201***	2.093	-8.204***	2.085
Student-teacher-ratio	-0.041***	0.009	-0.042***	0.009
Share of government funding	-0.140***	0.036	-0.143***	0.036
Admission by ability	2.383**	1.127	2.373**	1.126
School autonomy	3.734***	1.213	3.800***	1.203
Same textbook	-0.648	2.234	-0.741	2.263
School location:				
Small town (3,000-15,000)	0.257	3.143	0.284	3.131
Large town (15,000-100,000)	0.972	3.162	0.933	3.124
City (100,000-1,000,000)	3.887	3.861	3.954	3.853
Large city (¿1,000,000)	10.343**	4.111	10.418**	4.113
Country FE	Yes		Yes	
Country obs.	34		34	
School obs.	10,558		10,558	
Student obs.	249,968		249,968	
Avrg. R^2	0.49		0.49	

Notes: Dependent variable: PISA math score 2012. OLS regression weighted by students sampling probability. Cluster robust standard errors are given in parentheses. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

3.D Gender

	Male		Female	
Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.286	-3.454**	-0.098	-3.586**
	(0.690)	(1.741)	(0.644)	(1.467)
Ability Grouping × Comp		0.601*		0.665**
		(0.359)		(0.297)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,074	10,074	10,062	10,062
Student obs.	118,391	118,391	125,344	125,344
Avrg. R^2	0.48	0.48	0.50	0.50

Notes: Dependent variable: PISA math score 2012 of male (female) students. Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

3.E Variance

Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.113	0.120		
	(0.273)	(0.741)		
Ability Grouping × Comp		-0.045		
		(0.145)		
AG=1 (Some classes grouped)			0.611	-0.989
			(1.269)	(3.165)
AG=2			0.744	-0.097
			(1.245)	(2.345)
AG=3			0.596	0.587
			(1.042)	(2.869)
AG=4			-0.222	3.277
			(2.013)	(4.407)
AG=5 (All classes grouped)			-0.400	-1.975
			(1.475)	(3.717)
AG=1 × Comp				0.356
				(0.704)
AG=2 × Comp				0.173
				(0.468)
AG=3 × Comp				0.018
				(0.515)
AG=4 × Comp				-0.622
				(0.865)
AG=5 × Comp				0.301
				(0.774)
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,464	10,464	10,464	10,464
Avrg. R^2	0.34	0.34	0.34	0.34

Notes: Dependent variable: PISA 2012 standard deviation of math test scores per school. Least squares analysis using school weights. Robust standard errors are given in parentheses. Control variables: sd(age), share of females, shares of parents' education, sd(grade), share of grade repeaters, mean class size, sd(books at home), private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math-teachers, school location, admission by ability. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

3.F Robustness Checks

3.F.1 Science vs. Reading

	Science		Reading	
Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.020	-3.550***	-0.120	-3.304**
	(0.541)	(1.360)	(0.530)	(1.440)
Ability Grouping × Comp		0.671**		0.605**
		(0.273)		(0.286)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,558	10,558	10,558	10,558
Student obs.	249,968	249,968	249,968	249,968
Avrg. R^2	0.48	0.48	0.44	0.44

Notes: Dependent variable: PISA science score 2012 and PISA reading score 2012. Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

3.F.2 Alternative Definition of Group

Variables	(1)	(2)	(3)	(4)
Ability Grouping 2	-0.367	-5.052***		
	(0.972)	(1.883)		
Ability Grouping 2 × Comp		0.943**		
		(0.400)		
AG2=2 (Within-Class Grouping)			-4.518	-8.076
			(4.615)	(7.847)
AG2=3 (Some Between-Class Grouping)			-1.760	-18.126***
			(2.942)	(5.185)
AG2=4 (All Classes Grouped)			-1.542	-13.138**
			(3.085)	(5.747)
AG2=2 × Comp				0.785
				(1.893)
AG2=3 × Comp				3.469***
				(1.145)
AG2=4 × Comp				2.491**
				(1.233)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,563	10,563	10,563	10,563
Student obs.	250,042	250,042	250,042	250,042
Avrg. R^2	0.49	0.49	0.49	0.49

Notes: Dependent variable: PISA math score 2012. Reference category is AG2=0 (no within or between class grouping). Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

3.F.3 Immigrants

Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.296	-3.709***		
	(0.596)	(1.416)		
First generation	-7.929	-8.018	-7.647	-8.190
	(6.750)	(6.728)	(6.698)	(6.551)
Second generation	-4.288	-3.902	-4.007	-4.677
	(6.786)	(6.795)	(6.749)	(6.751)
Firstgen × Ability Grouping	0.147	0.173	0.079	0.339
	(2.530)	(2.547)	(2.493)	(2.479)
Secgen × Ability Grouping	0.235	0.083	0.144	0.418
	(2.409)	(2.426)	(2.380)	(2.394)
Ability Grouping × Comp		0.649**		
		(0.292)		
AG=1			-1.723	-11.033*
			(3.496)	(6.080)
AG=2			-1.695	-22.305***
			(2.639)	(5.769)
AG=3			0.279	-9.282
			(2.685)	(5.765)
AG=4			-2.151	-19.201**
			(3.258)	(8.193)
AG=5			-2.349	-19.417**
			(3.744)	(8.232)
AG=1 × Comp				2.037
				(1.516)
AG=2 × Comp				4.239***
				(1.209)
AG=3 × Comp				2.100*
				(1.223)
AG=4 × Comp				3.493**
				(1.737)
AG=5 × Comp				3.444**
				(1.656)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,580	10,580	10,580	10,580
Student obs.	273,720	273,720	273,720	273,720
Avrg. R^2	048	0.48	0.48	0.48

Notes: Dependent variable: PISA math score 2012. Reference category is AG=0 (no between class grouping). Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

3.F.4 Multilevel Model

Variables	(1)	(2)
Ability Grouping	-0.347	-4.686***
	(0.579)	(1.567)
Ability Grouping × Comp		0.833**
		(0.324)
Within-school SD	61.61	61.61
Between-school SD	32.47	32.47
Var. prop. attributed to schools (ρ)	0.21	0.22
Within-school var. prop. explained (%)	0.12	0.12
Between-school var. prop. explained (%)	0.70	0.70
Student controls	Yes	Yes
School controls	Yes	Yes
Country FE	Yes	Yes
Country obs.	34	34
School obs.	10,558	10,558
Student obs.	249,968	249,968
Avrg. R^2	0.48	0.48

Notes: Dependent variable: PISA math score 2012. Random effects regression weighted by students sampling probability. Standard errors are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

Chapter 4

An Experiment on Peer Effects under Different Relative Performance Feedback and Grouping Procedures[1]

Abstract

We conduct a laboratory experiment to test theoretical predictions from Thiemann (2017) on subjects' performance in an effort task conditional on their peer group's composition and relative performance feedback. Subjects are grouped either randomly or according to their ability, with the feedback being the maximum or average performance of their group. We are able to support theory-derived hypotheses on optimal performance and peer effects. While random grouping is beneficial for male subjects it is detrimental for female subjects. Evidence is found for output being more dispersed when the best group performance is given as feedback. Again we find gender differences with male subjects performing significantly better when they compare themselves to the best peer instead of the average, while the opposite is true for females.

JEL Codes: C91, J16, J24, M52

Keywords: Laboratory Experiment, Ability Grouping, Relative Performance Feedback, Peer Effects, Reference Dependent Preferences

[1]This chapter is co-authored by Niklas Wallmeier.

4.1 Introduction

In many areas of life, individuals that perform on a certain task find their performance evaluated relatively to that of other individuals of a reference group. In schools, students get feedback on their own grade, but usually they also receive some relative performance feedback within their class. In firms, employees can observe the performance of their team members, or, for instance, get feedback on the "best salesperson of the month". Assuming that individuals have reference dependent preferences (see e.g. Tversky and Kahneman, 1979; Clark and Oswald, 1998), with the relative performance feedback being the reference point, individual performance also depends on the kind of relative performance feedback and on the composition of the reference group.

The kind of relative performance feedback might vary between firms and institutions where individuals perform according to the firm philosophy. Some firms might actively highlight only the top performers in order to drive employees to perform better, some might not provide any feedback. Some teachers might provide students with the average class grade of the last exam as a reference point, others might provide their students with the full grade distribution. The reference point that is given can also vary with culture as acquired by groups of people that share a religion or ethnic origin. In more competitive cultures, for instance, individuals are expected to compete for the top positions. Accordingly the best performance is particularly emphasized and praised. In less competitive cultures, social comparison plays a less emphasized role and individuals are expected to conform to the average. The reference point does not need to be given exogenously by a firm or institution, but individuals might also endogenously choose whom to compare with according to their intrinsic values and culture. Accordingly, competitive individuals might choose a high reference point, for example the best student in class to compare to, whereas non-competitive individuals rather compare to the average.

Since the kind of relative performance feedback and the composition of the reference group can influence the effort provision of an individual, the question arises whether group composition and performance feedback can be optimized in order to maximize group performance. Thiemann (2017) deals with this question theoretically, focusing on a school environment and the question whether ability segregated classes (ability tracking or ability grouping) or classes with heterogeneous ability students (comprehensive schooling) are to be preferred. Theory predicts that it depends on the culture of competitiveness of the student body, i.e. on the kind of the reference point and the importance of social comparison.

The intuition of the hypotheses derived from this theory is that a high reference point

allows optimal incentivizing of individuals in groups of heterogeneous abilities. Here also subjects with very low ability are motivated by the top performers and exert effort in order to minimize the performance distance. In a system with ability grouping, where high-ability subjects are sorted into a high track and low-ability subjects into a low track, this positive effect is restricted to the individuals in the high track.

When an average reference point is given, the model predicts ability grouping to yield a higher average performance. This is driven by stronger motivation in a high-ability group due to the higher reference point compared to a heterogeneous group. This effect may on average outweigh the negative effect of ability grouping for low-ability students. Independent of the reference point the variance of performances increases under ability grouping, since ability grouping is always detrimental for low-ability subjects, but never detrimental for high-ability subjects.

The hypotheses from Thiemann (2017) are tested in a laboratory experiment. The controlled environment in the laboratory has several advantages: First, typical identification problems that arise in peer effects regressions (Manski, 1993; Moffitt et al., 2001) can be avoided by choosing a certain experimental design. Second, the channels that drive subjects to perform can be disentangled and an environment can be designed that closely matches the theoretical model.

In the experiment, subjects performed several periods of solving multiplication problems and earned a piece rate per correctly solved task. In the randomly-grouped treatment subjects were randomly sorted into groups of five. In the ability-grouped treatment groups were only composed either of high or low-ability subjects (within-subject design). After each period subjects were shown their own and either the best or the average group performance (between-subject design). In the analysis, we check for significant differences in mean performance under the different grouping and reference point regimes. Furthermore, regression analysis is used to test theory derived equilibrium performance and individual peer effects.

We find support for subjects behaving according to the theoretically derived optimal performance. However, further hypotheses on treatment differences, especially between random and ability grouping, cannot be confirmed. Still, regression analysis suggests that incentives differ conditional on whether the best or the average performance is available. These peer effects with respect to the reference point seem to be non-linear. Furthermore, gender differences in the reaction to different reference points and grouping procedures are evident.

The rest of the paper is organized as follows. Section 4.2 gives on overview of the related literature. Section 4.3 yields a brief recap of the theory by Thiemann (2017). The

experimental design employed to test this theory is described in Section 4.4. Results are presented in Section 4.5 including prima facie evidence, gender differences and regression analysis of peer effects. Section 4.6 concludes.

4.2 Related Literature

Our study contributes to two existing fields of economic literature. First, it is settled in the field of (laboratory) experimental evidence on peer effects. Second, it contributes to research that analyzes the effect of grouping procedures on performance, such as ability tracking/grouping or mixing/comprehensive schooling.

Recent literature has brought up strong evidence on the prevalence of peer effects in the lab. As Falk and Ichino (2006) show, pairing up workers can enhance productivity, independent of remunerations. Furthermore, they find that this effect is stronger for low productive workers. Hannan et al. (2008) highlight the influence of peer performance under diverse compensation schemes. They provide subjects with information on whether they performed above or below the 50th percentile of all participants. Their findings imply that the enhancing effect is present for individual compensation, but also that relative performance feedback can deteriorate performance when remuneration is based on peer output. When a tournament scheme is in place, especially loss-averse individuals are found to respond negatively to a rival's effort (Gill and Prowse, 2012). Also the subjects' particular rank apparently has influence on performance. Workers that are aware of a full ranking of peers' performances show 'first-place loving' and 'last-place loathing', i.e. subjects exert the highest increase in effort after being ranked first or last (Gill et al., 2015). Kuhnen and Tymula (2012) find the mere possibility of being evaluated relative to peers as potentially performance enhancing. That is, people work harder and expect to rank better when told that they may learn their ranking, relative to cases without feedback. When expectations concerning the rank have been exceeded, individuals decrease output but expect a better rank in the future, while the opposite is true for unfulfilled expectations. The influence of performance feedback has also been evaluated within network structures inside the lab (Beugnot et al., 2013). Here individuals get feedback on average performance of their neighbors in the network. Findings suggest that being surrounded by more productive peers increases individual performance. This positive peer effect is generally larger for male subjects. Our main contribution to these studies on peer effects in the lab is that we directly compare the influence of different types of performance feedback. Where past research only gives relative feedback on *average* peer achievement (Beugnot et al., 2013; Hannan et al., 2008; Azmat and Iriberri, 2010), we

contrast groups that receive feedback on average group performance with groups that receive feedback on the best peer performance.

These contributions from laboratory experiments are enriched by studies on field experiments that yield some first insights on the effects of group composition. Hamilton et al. (2003) find increased productivity in a garment manufacturing plant when group-based compensation is introduced. Here group composition also seems to matter, since more heterogeneous teams are more productive when average ability is held constant. In another study with cashiers in a supermarket evidence of positive productivity spillovers from highly productive workers is found, supporting the idea of heterogeneous group composition being favorable (Mas and Moretti, 2009)[2]. Peer effects have also often been analyzed in educational settings. Using a natural experiment, Azmat and Iriberri (2010) take advantage of university students receiving information on whether they perform above or below the class average as well as the distance from this average, for one year only. Providing this relative performance feedback, when individuals are rewarded according to their absolute performance level, leads to an increase of 5% in students' grades. Hoxby (2000) uses idiosyncratic variation (changes in gender and race composition of a class) to identify peer effects in a schooling environment. These are found to be significant and of notable size. If average peer achievement score increases by one point this leads to an estimated increase in individual student achievement by 0.15-0.4 points. Lavy et al. (2012) investigate whether such peer effects are driven by a specific sub-group of students. In particular, they find mainly evidence for extremely low-ability students having a negative impact on their peers' achievement.

Our study secondly contributes to the literature that directly addresses the effect of grouping individuals according to their ability ("ability tracking"). Effects of ability grouping can be due to mutual learning or norm setting within the group, of which the latter corresponds to the pure peer effect as analyzed in lab experiments. In a school environment the effects can also be due to different instructional pace. A number of field studies have analyzed the influence of ability tracking on student performance in school (a review is presented in Slavin (1990)). Effects of ability tracking on mean achievement are usually low and non-significant. In terms of performance distribution studies usually find that tracking harms low-ability students but benefits high-ability students (e.g. Argys et al., 1996; Hoffer, 1992). Of a different type is the recent study by Carrell et al. (2013), who conduct a field experiment with cohorts of entering freshmen at the United States Air Force Academy. They place half of the students into groups designed to maximize the academic achievement of the lowest third of the achievement distribution, i.e. these

[2]Van Veldhuizen et al. (2014) find no support for this property in the lab.

low-ability students are placed into squadrons with a high fraction of peers with high achievement scores. The results are, however, not as expected, since a negative and statistically significant treatment effect is observed for the lowest ability students. To our best knowledge there is no laboratory study on the effect of ability grouping on achievement. While the above mentioned field studies cannot disentangle whether different group compositions affect performance through mutual learning or through different group norms (Hamilton et al., 2003), our laboratory study can exclude mutual learning effects and focus on the latter.

4.3 Theory

The underlying theory is taken from Thiemann (2017) and shall be briefly summarized here. We assume that subjects in our experiment maximize utility by choosing an effort level. Assume that effort translates linearly into performance and that subjects have reference-dependent preferences as in Tversky and Kahneman (1979) with relative performance feedback being the reference point. Subjects face the following optimization problem:

$$\underset{p_i}{\text{Max}}\ u_i(p_i) = p_i + s \cdot v(p_i - r_i) - c(p_i, a) \tag{4.1}$$

$$\text{with}\quad v(p_i - r_i) = \begin{cases} \lambda \cdot (p_i - r_i) & if\ p_i < r_i \\ (p_i - r_i) & if\ p_i \geq r_i \end{cases} \tag{4.2}$$

$$\text{and}\quad c(p_i, a) = \frac{p_i^2}{2a} \tag{4.3}$$

Performance p_i is the number of correctly answered multiplication problems per period. Before each period, each subject is shown a reference point r_i, that yields information about the performance of the group members. Subjects' utility depends on a direct private component of utility and a comparison oriented component given by the value function $v(\cdot)$. In the experiment the direct private utility from performance is given because of direct remuneration of performance. The utility from the comparison oriented component is assumed to be larger the more competitive a subject is (s, with $s \geq 0$ is the degree of social comparison). For subjects performing below the reference point, the disutility from the difference to the reference performance is increasing with loss aversion, λ, with $\lambda > 1$. The cost of performance $c(p_i, a)$ increases in performance and decreases with ability a. A subject's optimal performance is then given by the following best response

function:[3]

$$BR_i(r_i) = \begin{cases} (1+\lambda s)a & if \quad p_i < r_i \\ (1+s)a & if \quad p_i > r_i \end{cases} \tag{4.4}$$

Optimal performance depends positively on ability a and competitiveness s. If the subject's performance is below the reference point, performance also depends positively on loss aversion (λ).

The derived best response function is the basis to compare equilibrium performances across different regimes. First, we compare performances for different reference points: the average performance among the other group members and the best performance among the other group members. Second, we compare a regime where subjects are randomly grouped with a regime, where subjects are grouped according to ability. In the latter we have groups consisting only of low-ability subjects and groups only with high-ability subjects. We follow the theoretical analysis of Thiemann (2017), where proof is found for four main hypotheses:

H1 *When the best reference point is given, average performance is higher under random grouping than under ability grouping.*

H2 *When the average reference point is given, average performance is higher under ability grouping than under random grouping.*

H3 *Low-ability individuals always lose under ability grouping.*

H4 *High-ability individuals gain from ability grouping when the average reference point is given, and are not affected when the best reference point is given.*

4.4 Experimental Design

4.4.1 Effort Task

The main component of the experiment is a real-effort task, that subjects were asked to perform on a computer in the laboratory. We take the idea for this particular effort task from Dohmen and Falk (2011). Subjects were asked to solve as many multiplication problems as possible in five periods of four minutes each. In particular we asked subjects to multiply one-digit numbers (3-9) with two-digit numbers (11-99). By remunerating subjects with a piece rate per solved problem, they were linearly incentivized. Every

[3]For simplification we ignore the case where $p_i = r_i$. See Thiemann (2017) for the full solution.

subject was given the same problems in the same order to ensure that the difficulty of the problems was the same for everyone. Problems were not randomly created, but purposefully designed such that the difficulty of problems would vary to the same extent within each period. In case subjects answered a problem incorrectly the screen reported "false" and subjects had to repeat it. This was implemented to avoid that subjects search for easy problems. The number of correctly answered problems in each period was always shown at the top of the input screen. Subjects were instructed not to use any helping devices such as calculators, mobile telephones or paper and pencils (see instructions in Appendix 4.A). An example of the input screen showing the multiplication task is given in Appendix 4.B.

Multiplication problems were chosen as an effort task to ensure that performances during the experiment depend both on ability and effort. On the one hand the given task is a good proxy for cognitive abilities and generates performances heterogeneously enough in order to do ability specific analysis and to group subjects on that basis. On the other hand the task offers sufficient scope to vary effort, since solving the problems needs high concentration and is thus costly. We expect to find learning effects during the experiment as did Beugnot et al. (2013), but they are expected to be small (Roth, 2001).

4.4.2 Treatments

In order to test the hypotheses ***H1*** and ***H2*** we implement a two-by-two design to compare mean group performances along the two major treatments: *best* vs. *average* reference point and *ability grouping* vs. *random grouping*. To test hypotheses ***H3*** and ***H4*** we do not need additional treatments, but can compare low and high-ability subjects between these four main groups. In addition, we have a baseline treatment that is used to group subjects according to ability. Subsequent to the experiment we measure individual loss aversion and competitiveness by survey questions in order to test the theoretical optimal performance.

(a) *Baseline Treatment*

All subjects participated in the baseline treatment, which is the first period of 4 minutes solving multiplication problems. In this period subjects did not receive any information on reference points and were not sorted into groups. They only received information on their own performance, i.e. they could see their number of correctly answered multiplication problems after the period.

(b) *Best vs. Average Treatment*

The *best* vs. *average* treatments are modeled in a between-subject design, i.e. subjects

are either shown the *best* reference point or the *average* throughout the session. This is done to avoid a demand effect that would probably arise, if subjects are offered two different reference points subsequently. During the experiment subjects are sorted into groups of five. These groups serve the only purpose of providing the reference point for each subject. In the *best* treatment we provide subjects with information on the *best* performance of their group after every multiplication period. If the subject herself had the best performance we gave information on the second best performance. The subjects from the *average* treatment were given information about the *average* performance of their group, excluding the subject's own performance.

(c) *Ability Grouped vs. Randomly Grouped Treatment*

The grouping treatments are modeled in a within-subject design. All subjects went through two periods of the *randomly grouped* treatment and through two periods of the *ability grouped* treatment. In half of the sessions the *randomly grouped* treatment was conducted before the *ability grouped* treatment, in the other sessions the other way around. A within-subject design was chosen to mimic a real life situation, e.g. schools in which students first go through comprehensive primary schooling and are then grouped into tracks according to ability in secondary or tertiary schools. Also, the demand effect is supposedly weak in this treatment which is why a between-subject design could be chosen to increase power and statistical relevance without disadvantages. In the *randomly grouped* treatment subjects were randomly grouped with other subjects. This resulted in groups of subjects with heterogeneous abilities. For the *ability grouped* treatment subjects were ranked according to their performance in the first period (*baseline*). All subjects that performed in the top 50% of the ability distribution were sorted into a high track (high-ability type subjects), and those that performed in the bottom 50% were sorted into a low track (low-ability type subjects). Groups under the *ability grouped* treatment were then only randomly composed of subjects within these tracks. This resulted in groups of subjects with rather homogenous abilities, half of the groups with low-ability subjects and half with high-ability subjects.

Table 4.1 illustrates the composition of the sessions with respect to the reference point and the ordering of the grouping procedure. Altering the two possible reference point frameworks and switching the order of the grouping treatments allows observing all four possible setups. The crossover design with respect to the grouping treatments has two crucial advantages. First, we are able to deal with potential order effects. Practically that means, biases from being grouped by ability first and randomly later and vice versa can be excluded. In addition, we can also account for potential learning effects. Since we cannot rule out in advance that the participants' capability in solving multiplication tasks

Table 4.1: Session Designs

Reference Point: Average			**Reference Point: Best**		
(1)			(2)		
baseline →	random grouping →	ability grouping	baseline →	ability grouping →	random grouping
(1 period)	(2 periods)	(2 periods)	(1 period)	(2 periods)	(2 periods)
(3)			(4)		
baseline →	ability grouping →	random grouping	baseline →	random grouping →	ability grouping
(1 period)	(2 periods)	(2 periods)	(1 period)	(2 periods)	(2 periods)

increases due to learning or practice over time, we need to have the treatments played in each stage with equal frequency. Therefore, this design enables to disentangle learning from the treatment effects.

4.4.3 Experimental Procedure

The procedural details of a particular session are summarized in the following. In advance, the participants were told about the entire procedure of the experiment in a written overview (see instructions in Appendix 4.A). In addition, they received detailed information on the screen before a particular treatment started. This way, we ensured that the subjects were always aware of the character of the information they received. In the instructions we emphasized that their behavior in the experiment would remain anonymous. In particular, subjects were also told that they were going to be tracked according to their performance in the first period (*baseline*). This was done at the beginning to ensure that all subjects, no matter in which session, had the same information at the start of the experiment. Nevertheless, subjects had no incentives to behave strategically in the baseline period, since monetary incentives were the same across all periods no matter what track the subject was in. After giving the instructions, the subjects participated in a trial period of the multiplication task (30 seconds) to make themselves familiar with the displayed screens and the answer tool. When all participants had finished this period and no further clarifications were requested, the actual experiment started with a single period of the *baseline* treatment. Based on their performance in that period, the subjects were sorted into a low or a high track. This sorting decision was displayed to the subjects before the two periods of *ability grouped*. Subsequent to *baseline* two periods of *randomly grouped* [*ability grouped*] were played. After finishing the second period of *randomly grouped* [*ability grouped*], the experiment proceeded with two more periods of the

ability grouped [*randomly grouped*] treatment. This time, groups were shuffled within the same track [within the whole subject pool]. One period always consisted of the following four parts. First, groups of 5 members were formed by reshuffling the subjects randomly [within their track]. Second, the subjects were informed about their group members' performance (*best* or *average* reference point) of the previous period and also reminded of their own past performance. Third, the multiplication task was played for four minutes. And finally, a screen displayed feedback on the reference point and own performance in the just played multiplication task.

At the end of the experiment, one of the five periods was picked randomly as payout period and the profit was displayed to each subject. Finally subjects were asked to complete a questionnaire, in which we asked for socio-economic and demographic variables (e.g. subjects' gender, age, and field of study) as well as their competitiveness and loss aversion (see Appendix 4.C for questionnaire). To construct the variable about competitiveness we asked subjects to decide between a game where they could earn a piece rate or the same game where they could earn a prize by beating a component. The coefficient of loss aversion was elicited by a method developed by Abdellaoui et al. (2008). Subjects were asked to accept or reject different types of lotteries to elicit their certainty equivalents for losses and gains (see Appendix 4.C.1).

The experiment was programmed with zTree (Fischbacher, 2007) and conducted at the experimental laboratory of the University of Hamburg in June and July 2015. We used hroot for recruitment (Bock et al., 2014). We ran four sessions with a total of 120 participants. The subjects were students of the University of Hamburg of which 58 were female and 62 male, with an average age of about 25 years. One correct answer in the relevant periods was exchanged for 30 Euro cent. On average, a participant received a payout of 14 euros, including the show up fee of 5 euros. The sessions took about 60 minutes each.

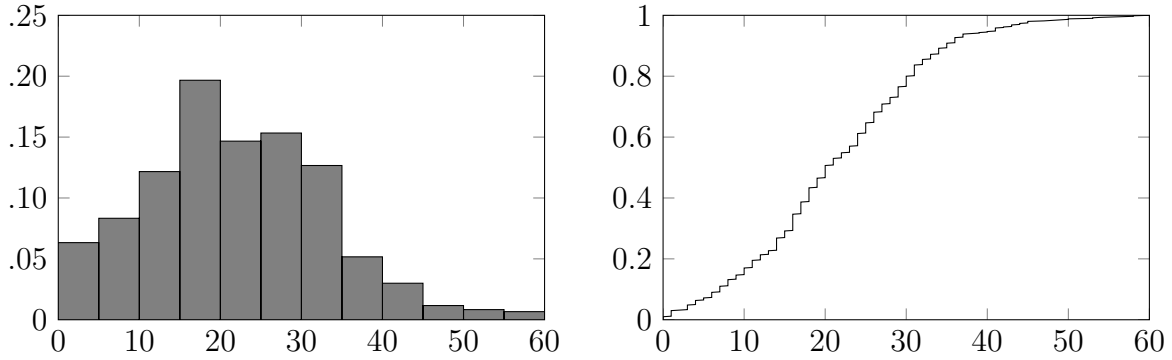
4.5 Results

4.5.1 Summary statistics and prima facie evidence

In this Section, we give an overview of the results by highlighting the data on the aggregate level. Performance is characterized under different grouping regimes and reference point settings. We are also able to test theoretical predictions on aggregate outcome, before the analysis focuses on the individual level.

Looking at the distribution of output over the entire experiment, we see that the

subjects performed sufficiently heterogeneous in the effort task. More precisely, output has a range from no correct answer up to a total of 60 correctly solved multiplications with a mean of 21.4. It can be taken from Figure 4.1 that performance is positively skewed around the median of 20, while a Shapiro-Wilk test rejects normality of the data ($z < 0.001$).



(a) Histogram (b) CDF

Figure 4.1: Distribution of Correct Answers

Figure 4.2 illustrates mean performance and its standard deviation per period to see how performance evolved over the periods in all treatments. There is a slight tendency of mean performance to increase steadily (from 18.3 to 24.1, dark gray bars), indicating that subjects improve over time independently of the treatment. For treatment comparisons, however, this should not matter, since it is controlled for the time effect by the crossover design. Further, evaluating learning separately for high-ability subjects (light gray bars) and low-ability subjects (white bars), suggests that the improvement is similar for both types. We find a gap of on average 13.3 correct answers between subjects with an above-median output and those who performed below-median in the first round. This difference remains fairly constant throughout the periods and stays in the range between 13.3 and 15.8. This indicates that the performance of the two types neither disperses nor convergences to a common level. This learning effect being constant across abilities facilitates the analysis on the individual level since it can be controlled for by the inclusion of period fixed effects.

In a next step, we take a look at potential differences in performance induced by *i) the reference point setting* and *ii) the grouping procedure* (see Figure 4.3). When the distribution is decomposed with respect to the given reference point (*best vs. average* treatment), we find differences in the maximum performance across both grouping procedures. With the 95th percentile at a level of 44, the *best* treatment shows higher peak output compared

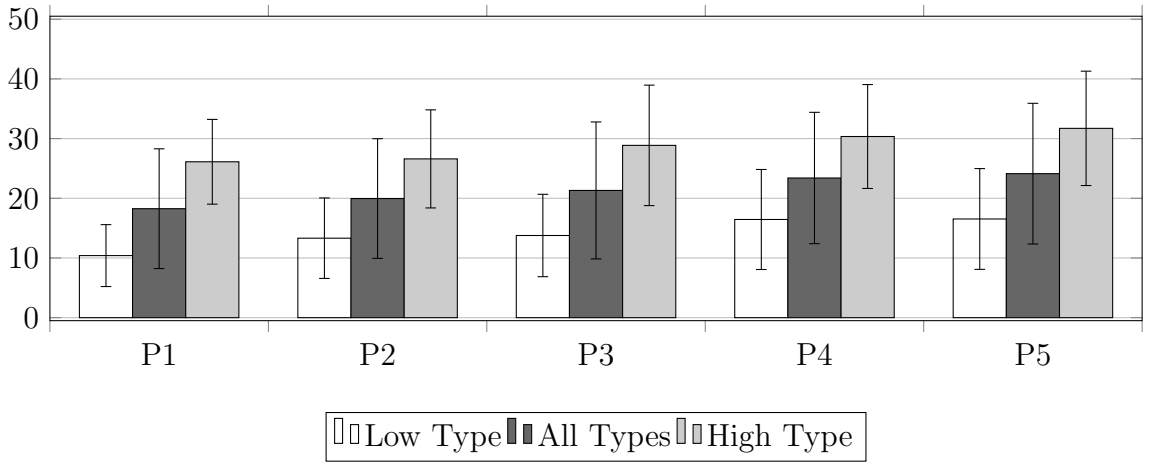
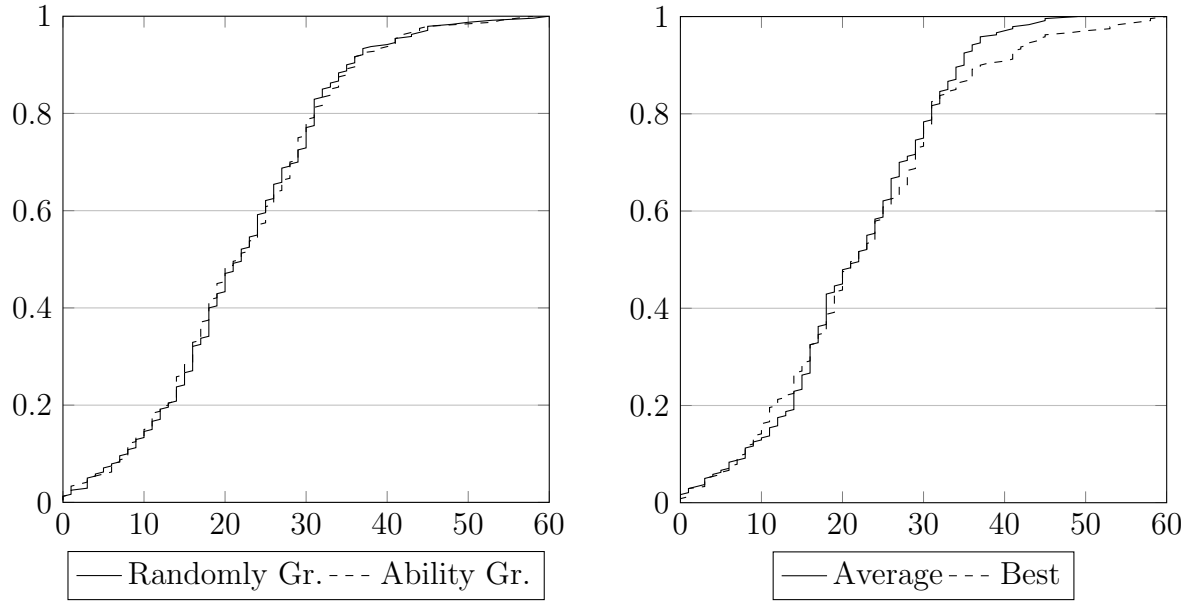


Figure 4.2: Mean Performance across Periods

to 37 under the *average* setting. A Brown and Forsythe test confirms that the variance of performance under the *best* treatment is significantly different (p = 0.045). On the other hand, we find no differences in peak output for the different grouping procedures (for both the 95th percentiles at 41). When we compare the settings with respect to mean differences, there is neither a statistically significant difference between the reference points (*best*: 22.67, *average*: 21.73, Mann-Whitney-U test (MWU): $p = 0.795$), nor between the two grouping treatments (*randomly grouped*: 22.23, *ability grouped*: 22.18, Wilcoxon signed-ranks test (WSR): $p = 0.807$). Since we observe each individual twice in a treatment, we use the average of a subject over the two periods as an observational unit for the MWU and WSR tests.

We continue with an analysis of the theoretical predictions from Section 4.3. To verify hypotheses ***H1*** and ***H2*** we contrast the mean performance of the two grouping scenarios under a given reference point. Figure 4.4 displays the mean outcome and standard deviation for both random (RG) and ability grouping (AG) given average group performance as reference point (AVRG) on the left-hand side, and the best group performance as reference point (BEST) on the right-hand side. Evaluating performance of all subjects (dark gray bars) under the *best* setting suggests that our experiment cannot confirm hypothesis ***H1***. Both random and ability grouping yield a mean performance of 22.7. Also with respect to the *average* setting we have to reject hypothesis ***H2***, since the output under both grouping treatments is not significantly different (21.8 vs. 21.7, WSR: $p = 0.593$).

To test hypotheses ***H3*** and ***H4***, we compare the mean performances separately for high-ability subjects (light gray bars in Figure 4.4) and low-ability subjects (white bars) across the grouping and reference point treatments. Hypothesis ***H3*** predicts a lower mean for low-ability subjects in an ability grouped setting compared to random grouping, irre-



(a) Performance per Grouping Treatment (b) Performance per Reference Point

Figure 4.3: Cumulative Distribution of Performance per Treatments

spective of the reference point. While this cannot be found for the *best* setting (*randomly grouped*, 15.15; ability tracking, 15.20), the *average* setting produces a lower mean output for ability grouping, although it is not statistically significant (15.15 vs.14.67, WSR: $p = 0.275$).

According to hypothesis ***H4***, we expect an output-enhancing treatment effect from ability grouping for high-ability subjects when the average group performance is given as a reference point. As it can be seen from Figure 4.4, the mean performance of high-ability subjects in the *average* setting is not significantly different across the grouping treatments (*randomly grouped* 28.77; *ability grouped* 28.42, WSR: $p = 0.750$).

Summarizing the analysis of aggregate treatment effects, we find some evidence for an impact of different reference points on subjects' performance with the *best* reference point inducing higher peak output. Grouping procedures, on the other hand, do not yield significant differences in aggregated performance. Therefore, we are not able to support our theoretical predictions on the aggregate level. Possible explanations are found in differing reactions to the grouping procedures by gender. In Section 4.5.2 we find evidence for opposing effects for men and women that cancel each other out on the aggregate. Furthermore, the results for the grouping procedures might be explained by non-linear reactions to the reference point, precisely, by diminishing sensitivity. If the motivating effect diminishes with the distance from the reference point and becomes small for a sufficiently large share of subjects, we might not observe the treatment effects on the entire

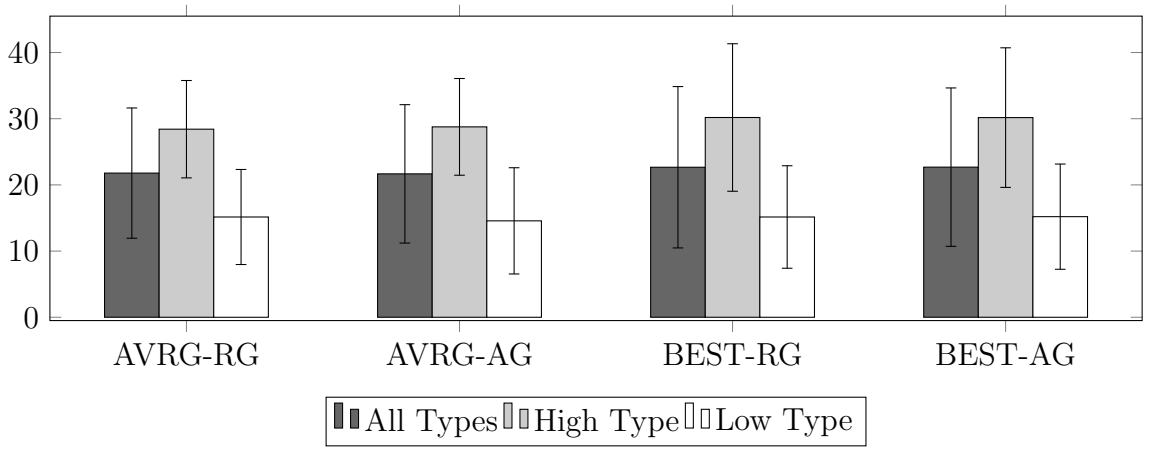


Figure 4.4: Performance per Reference Point and Grouping Treatment

sample, even if they are present for the individual subject. Consequently, we investigate how varying environments influence performance on the individual level with the help of regression analysis in Sections 4.5.3-4.5.5.

4.5.2 Gender Differences

Possible variations in performance according to gender might be a reason for missing differences at the aggregate level. The existence of gender differences in competitiveness is well documented. This research generally finds that men perform better in competitive environments (e.g. tournaments), whereas women's performance does not change in a tournament-based compensation scheme compared to a piece rate (Gneezy et al. (2003), Niederle and Vesterlund (2007)). In our setting there is no tournament, where the best performer receives a monetary prize, but the fact that we show the best performance to the subjects might be incentive enough to evoke similar effects. We thus expect men to perform better in the *best* treatment than in the *average* treatment, whereas female performance should stay the same.

Figure 4.5 shows the differences between male and female subjects in the *best* and the *average* treatment. Comparing mean performance in the *average* treatment first, it appears to be almost the same for men and women (women: 22,4, men: 21,3). In the *best* treatment, however, male subjects perform on average significantly better than women (women: 19.7, men: 26.5, MWU: p = 0.005). Men also perform significantly better in the *best* treatment than in the *average* treatment (MWU: p = 0.044), while female subjects perform worse in the *best* treatment compared to the *average* treatment (MWU: p = 0.111). Further, it can be seen that female performance has a lower variance than male performance across both treatments. A Brown and Forsythe test confirms that

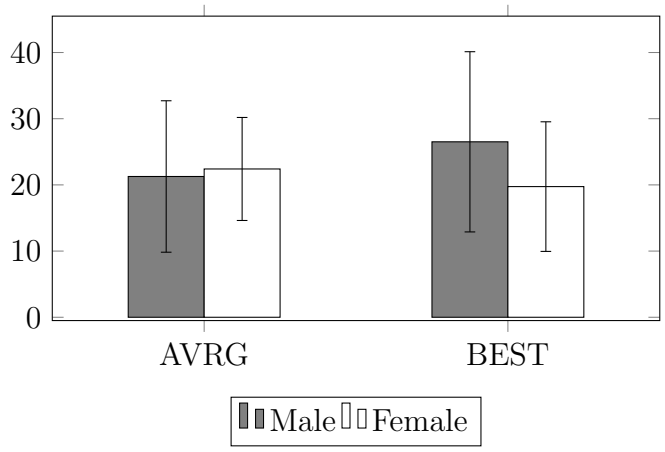


Figure 4.5: Average Performance by Reference Point Treatment and Gender

the variance of male and female performance is significantly different, overall and within the *best* and *average* treatment (all with $p = 0.000$).

Finally, we are interested in the influence of the two grouping regimes on average performance of male and female subjects. Figure 4.6 suggests that both grouping procedures affect the performance of a subject in the same way independently of the type, but differently for the gender. Both high and low types of female subjects benefit from *ability grouping*. Overall we find a significant difference between the two grouping procedures for women (AG: 21.5, RG: 20.2,WST: $p = 0.074$). The opposite is true for male subjects, who on average perform significantly better under *random grouping* (AG: 22.8, RG: 24.1, WST: $p = 0.031$). Taking into account that men are more competitive than women, this result can be interpreted as a confirmation of the theoretical prediction from Thiemann (2017) that random grouping is beneficial when subjects have a competitive mindset and

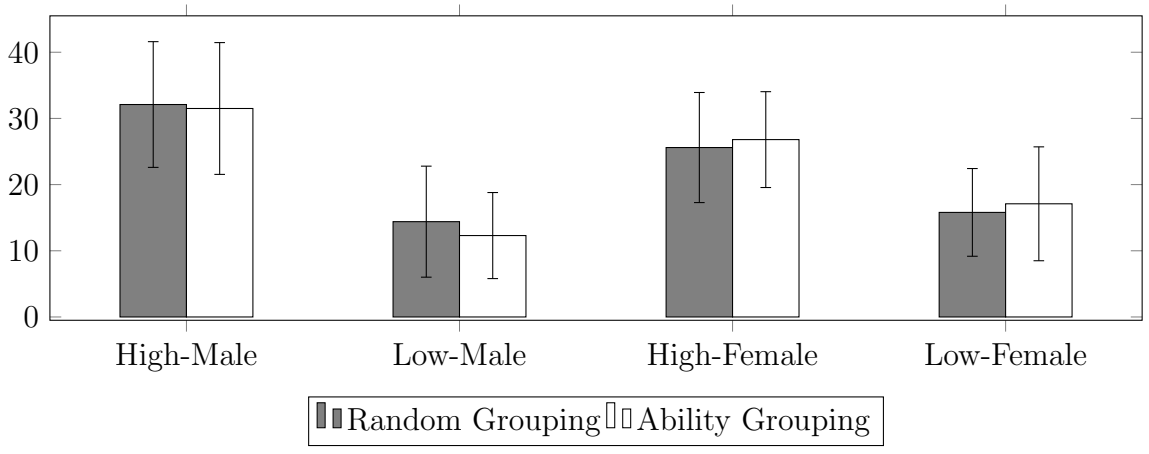


Figure 4.6: Average Performance under Random and Ability Grouping by Gender and Type

detrimental when subjects are non-competitive.[4]

4.5.3 Testing Optimal Performance

The hypotheses tested in Section 4.5.1 were derived from the optimal performance as theoretically derived in Section 4.3. Whether individual subjects behave according to the derived best response function can be tested directly in a system of regressions. If subjects behave optimally, their performance should depend positively on ability (a_i) and competitiveness ($comp_i$). If the subject's performance is below the reference point, performance should also increase with the degree of loss aversion ($lossavers_i$).

$$p_{it} = \alpha + \beta_1\ lossavers_i + \beta_2\ comp_i + \beta_3\ a_i + \mu_t + \epsilon_{it} \quad if \quad p_{i,t-1} < r_{i,t-1} \tag{4.5}$$

$$p_{it} = \alpha + \beta_1\ lossavers_i + \beta_2\ comp_i + \beta_3\ a_i + \mu_t + \epsilon_{it} \quad if \quad p_{t-1} > r_{t-1} \tag{4.6}$$

The dependent variable is performance of subject i in period t. The first regression only includes subjects that performed below the average (or best) performance of their current group members in the last period. The second regression includes those that performed above. The three covariates of interest are derived from questions that subjects answered in the questionnaire subsequent to the experiment. Accordingly close to 27.5% of the participants are categorized as competitive, since they opted for the tournament in the question. Estimated coefficients of loss aversion had a mean of 3 and a standard deviation of about 3.5. As a control for ability we asked subjects for their last math grade at school (ranging from 1-6, with 1 being the best grade). We use math grades as a control for ability instead of baseline performance, since the latter cannot be considered an objective measure. Subjects go through the baseline treatment knowing from the instructions, that they will get relative performance feedback later and that they will be tracked according to their performance in this period. Because of this potential lead effect we chose to use a measure for ability that was determined outside the experiment. Math grade should be a valid control for ability since the multiplication task requires basic mathematical abilities that were taught and regularly used at school.

The regression also includes period and session dummies (μ_t) to control for period and session specific effects, especially for learning effects. We expect β_1, β_2 and β_3 to be positive and significant in Equation (4.5) and only β_2 and β_3 to be positive and significant

[4] In the survey after the experiment we included a question in which participants had to choose between a tournament-based compensation scheme and a piece rat. 24% of the female subjects chose the tournament and 30% of male subjects.

in Equation (4.6). Results of Ordinary Least Squares (OLS) regressions with standard errors clustered at the individual level to control for serial correlation in the error term are reported in Table 4.2, separately for the *best* and the *average* treatment.

Table 4.2: Testing Theory Derived Optimal Performance

Variables	Average		Best	
	Below (1)	Above (2)	Below (3)	Above (4)
Loss Aversion	0.458**	0.138	0.980**	-0.418
	(0.180)	(0.166)	(0.415)	(0.691)
Competitiveness	-3.145	3.859	-0.995	9.896*
	(3.162)	(2.600)	(3.392)	(5.534)
Math Grade	-2.084**	-1.346	-0.185	-7.832***
	(0.874)	(1.248)	(1.288)	(1.684)
Constant	20.630***	25.125***	18.343***	44.431***
	(3.416)	(3.763)	(4.722)	(6.595)
Period FE	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes
R^2	0.22	0.20	0.13	0.56
Adj. R^2	0.15	0.13	0.08	0.46
N	85	91	130	38

Notes: Ordinary least squares regressions. Dependent variable: Number of correct answers. Regressions include periods 2-5. Robust standard errors in paranthesis are clustered at the individual level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

As predicted the coefficient of loss aversion has a positive and significant impact on performance only for subjects whose past performance was below the reference point both in the *best* and the *average* treatment. Precisely, for subjects below the reference point in the *average* treatment an increase of the coefficient of loss aversion by 1 induces on average an increase of the correctly answered multiplication tasks by roughly 0.5 and almost by 1 in the *best* treatment. The indicator for competitiveness has no positive impact on performance. The ability control strongly predicts performance of the top subjects in the *best* treatment (see high R^2) and for below average subjects in the *average* treatment, but largely fails to explain performance of the other subjects. Taken altogether, especially the estimates for loss aversion that drives performance below the reference point can confirm the theoretical prediction.

4.5.4 Linear Peer Effects

In the preceding Section we have shown that performance increases in loss aversion if the subject's performance is below the reference point. Here we estimate the size of the

average effect of the reference point on performance. Typically these *peer effects* are empirically modeled by the linear-in-means-model, meaning that performance of a single subject is regressed on the average performance of the subjects' reference group (see e.g. Brock and Durlauf (2001)). We proceed in this way for the *average* treatment, while for the *best* treatment we regress individual performance on the best performance of each group. The following regression with period fixed effects μ_t and covariates $\mathbf{X}_i$ is estimated separately for the *best* and *average* treatment.

$$p_{it} = \alpha + \beta \; refpoint_{it} + \mathbf{X}_i\gamma + \mu_t + \epsilon_i \tag{4.7}$$

The variable $refpoint$ is the average (best) performance of the current group members from the last period that was shown to the subjects before each multiplication period. Usually two identification problems in peer effects regression arise: self-selection (also: correlated effects) and the "reflection problem" (also: contextual effects) (see Manski, 1993). The first, self-selection of individuals into groups, does not arise in an experimental setting, since subjects are randomly allocated to groups and here also reshuffled after every period. The second, the "reflection problem" arises when individuals interact in groups that include themselves and the peer effect is correlated to mean characteristics of the group. In our setting this problem does not arise since subjects are only shown a reference point that does not include their own achievement and is thus exogenous (they are shown the *average* or the *best* achievement among the *other* group members). Also subjects are not given any information about the characteristics of their group members apart from their average or best performance.

If performance below the reference point increases linearly in loss aversion, the size of the peer effect should be larger in the *best* treatment than in the *average* treatment. The way in which subjects react to a reference point should strongly depend on subject specific characteristics, as suggested by theory e.g. on factors like loss aversion, competitiveness and ability. These factors again might vary, for instance, with the cultural background or the gender of the individual subject. Thus, we estimate a model that only includes $refpoint$ as a first step. The estimated coefficient gives the total impact of the reference point on performance, including any effect that might work through different subject characteristics such as culture, gender or ability. In a second step we include control variables for subject background factors gathered in the questionnaire subsequent to the experiment to see how this changes the impact of the reference point (these are:

female, years since Abitur[5], studies math[6], income[7]). To analyze which factors drive the sensitivity to the reference point, we include some interactions of $refpoint$ with subject characteristics in a third step. We use an OLS approach with clustered standard errors at the individual level. We expect β to be positive in specifications (1), (2), (4) and (5).

Table 4.3: Linear Peer Effects

Variables	Average			Best		
	(1)	(2)	(3)	(4)	(5)	(6)
Reference Point	0.569***	0.475***	0.749**	0.298***	0.216**	0.211
	(0.115)	(0.109)	(0.350)	(0.079)	(0.087)	(0.324)
Math Grade		-2.417***	1.140		-2.662**	-2.868
		(0.816)	(1.571)		(1.181)	(1.813)
Female		-2.309	2.412		-7.178***	-10.395*
		(2.408)	(4.704)		(2.624)	(5.915)
Years since Abitur		-0.601	-2.029***		0.147	0.012
		(0.412)	(0.602)		(0.246)	(0.328)
Studies Math		1.353	6.351		9.086***	9.897
		(2.014)	(4.365)		(2.695)	(6.208)
Income		1.055	1.134		0.300	1.355
		(0.746)	(1.212)		(1.016)	(2.413)
Reference Point × Math Grade			-0.189***			0.005
			(0.061)			(0.047)
Reference Point × Female			-0.248			0.101
			(0.177)			(0.150)
Reference Point × Years since Abitur			0.079***			0.005
			(0.029)			(0.013)
Reference Point × Studies Math			-0.237			-0.024
			(0.176)			(0.161)
Reference Point × Income			-0.005			-0.033
			(0.045)			(0.065)
Constant	9.629***	18.827***	12.772	11.232***	23.118***	23.415*
	(2.625)	(4.995)	(9.159)	(2.546)	(6.623)	(12.147)
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.16	0.31	0.37	0.10	0.35	0.36
Adj. R^2	0.14	0.28	0.32	0.08	0.32	0.31
N	240	236	236	240	228	228

Notes: Dependent variable: Number of correct answers per period. Robust standard errors in paranthesis are clustered at the individual level. Regressions include period 2-5. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

[5] *Abitur* is the name of the diploma awarded to students at the end of secondary schooling in Germany.

[6] The variable *studies math* is a dummy that takes on the value 1 if the subject studies a course that includes mathematics as a major component, such as information systems, economics, business, physics or mathematics.

[7] The variable *income* is an ordered categorical variable taking on the following values of disposable income per months (in Euros): 1 = less than 400, 2 = 400-600, 3 = 600-800, 4 = 800-1000, 5 = 1000-1200, 6 = more than 1200.

From the results reported in Table 4.3 we see that in both treatments individual performance increases in the reference point. However, the effect is almost twice as large in the *average* treatment. When the reference point is one correct answer higher, individual performance increases on average by more than half a correct answer in the *average* treatment and only by 0.3 correct answers in the *best* treatment. In both treatments the impact of the reference point decreases once we control for subject characteristics, but it remains positive and significant. Including interactions does not shed any light on what drives the sensitivity to the reference points in the *best* treatment. In the *average* treatment, however, we find that subjects that have a better math grade and older subjects (subjects whose graduation from school is longer ago) react more strongly to the reference point. Unlike Beugnot et al. (2013) we find no difference in the reaction to reference points between male and female subjects. This effect might be taken up by the math grade, which is significantly better for female subjects (pairwise correlation: -0.128***).

Experimental evidence suggests that the importance of the reference point diminishes once a certain performance hierarchy has been established (Kuhnen and Tymula, 2012). Indeed, regressions including both the average and best treatment (see Table 4.4) show that the impact of the reference point decreases considerably after the third period and is only significant at the 10% level in the fifth period.

Table 4.4: Diminishing Importance of Reference Point

Variables	2nd Period	3rd Period	4th Period	5th Period
Reference Point	0.437***	0.574***	0.285***	0.200*
	(0.084)	(0.111)	(0.107)	(0.115)
Constant	11.949***	8.803***	16.636***	18.744***
	(2.103)	(2.879)	(2.932)	(3.219)
Session FE	Yes	Yes	Yes	Yes
R^2	0.17	0.20	0.06	0.02
N	120	120	120	120

Notes: Ordinary least squares regression. Dependent variable: Number of correct answers in the indicated period. Robust standard errors in paranthesis are clustered at the individual level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

4.5.5 Non-linear Peer Effects

So far, we looked at average peer effects. However, this approach might not capture the theoretical prediction of a non-linear reaction to the reference point with the effect being larger below the reference point due to loss aversion. Also, unlike suggested by theory we have seen in the last Section that an average reference point has a higher impact on

individual performance than the best reference point. A reason for this could be non-linear effects and diminishing sensitivity with respect to the reference point as suggested by Tversky and Kahneman (1979). The motivating effect of the reference point might become smaller the further away a subjects' performance is from the reference point. To find the effect of the distance to the reference point in our sample we use a differencing method, i.e. the dependent variable is the change in correctly answered problems compared to the period before. With this approach we can avoid multicollinearity of the subjects' performance and the distance to the reference point. We can also eliminate time-invariant factors like subject ability and concentrate on what causes the change in performance between periods. The following regression is estimated separately for the *best* and *average* treatment:

$$\begin{aligned}\Delta p_{it} =& \alpha + \beta_1 below_{it-1} + \beta_2 absdist_{it-1} + \beta_3 absdist_{it-1} \times below_{it-1} \\ &+ \beta_4 trackdec_{it} + \beta_5 trackdec_{it} \times lowtype_i + \mu_t + \mu_i + \Delta\epsilon_{it}\end{aligned} \tag{4.8}$$

The variable *absdist* is the absolute distance in points of the subjects last period performance to the reference point, both of which is shown to subjects at the beginning of each period. The variable *below* indicates whether the subject had performed below the reference point in the last period. The only other thing that changes with t is that subjects are told before the *ability grouped* treatment whether they were sorted into the low or high track. To control for this we include a dummy for the period in which subjects received this information (*trackdec*). We also include an interaction of *trackdec* with *lowtype*, which indicates whether subjects were sorted into the low track. At the cost of explanatory power, we estimate fixed effects models with subject and period fixed effects to eliminate biases due to unobserved subject characteristics and learning effects.

To find proof of a peer effect that is larger below the reference point, we would expect $\beta_1 > 0$. In order to find proof of diminishing sensitivity as suggested by Tversky and Kahneman (1979), we would expect $\beta_2 < 0$ and $\beta_2 + \beta_3 < 0$, i.e. both, above and below the reference point, it holds that increasing distance to the reference point lowers the enhancing effect on performance. Results are reported in Table 4.5.

Specification (1) shows that, while there is no increase in performance for subjects above the reference point (see constant), subjects who were told that they performed below the average improve their number of correctly answered questions by more than four in the following period. In contrast, no significant difference can be found for the *best* treatment. Since the output of those below the average performance is on average clearly lower than of those who only failed to make the top position, this result suggests

Table 4.5: Effect of Distance to Reference Point

Variables	Average		Best	
	FE (1)	FE (2)	FE (3)	FE (4)
Below the Reference Point	4.381***	1.810	0.409	0.549
	(0.861)	(1.315)	(0.671)	(1.729)
Absolute Distance to the Reference Point		-0.261**		-0.180
		(0.110)		(0.181)
Absolute Distance to the Reference Point × Below the Reference Point		0.505***		0.359*
		(0.176)		(0.190)
Period of Tracking Decision	-0.089	-1.980*	-0.340	-2.950**
	(0.902)	(1.193)	(1.092)	(1.411)
Period of Tracking Decision × Low Type		3.138**		6.097***
		(1.535)		(1.898)
Constant	-0.046	1.457	1.243	-0.920
	(0.913)	(1.197)	(1.207)	(1.667)
Period FE	Yes	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes	Yes
R^2	0.17	0.22	0.01	0.11
Adj. R^2	-0.13	-0.08		-0.23
N	240	240	240	240

Notes: Dependent variable: Change in performance compared to last period. Standard errors in paranthesis. Regressions include periods 2-5. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

that the effect not only depends on being below the reference, but also on the size of the gap. Including the variable on the distance, specification (2) shows that for the *average* treatment there is evidence for diminishing sensitivity above the reference point, but for increasing sensitivity below the reference point. Also in the *best* treatment, where almost every subjects is below the reference point, we find weak evidence for increasing sensitivity with growing distance to the best performance (see specification (4)).

Furthermore, evaluating the output subsequent to the tracking information, we find patterns that also have been found in previous literature (Kuhnen and Tymula, 2012). Subjects that are told that they were sorted into the low track do significantly improve in the following period especially in the *best* treatment. Here they solve on average three tasks more than in the previous period. Opposite, those who are told that they are sorted into the high track do rather deteriorate. Adjusting the performance upwards after being evaluated in the bottom tier is also consistent with the finding of 'last-place loathing' behavior by Gill et al. (2015).

4.6 Conclusion

In this paper we conducted an experiment to test the theoretical prediction that subjects behave differently in an effort task according to their peer group's composition when given different relative performance feedback (Thiemann, 2017). While support is found for subjects behaving according to the theoretically derived optimal performance, further hypotheses on treatment differences, especially between random and ability grouping, cannot be confirmed. However, a robust finding suggests that incentives created by showing the best performance as a reference point, differ from those evoked by showing an average reference point. Hence, performances in the best treatment are more dispersed, with outliers at the top of the distribution. Especially male subjects perform significantly better when they are offered the best performance of their group as relative performance feedback instead of the average. With respect to the grouping treatments there is no significant difference on an aggregate level, but once we distinguish between male and female subjects we find a significant negative effect of random grouping for female subjects and a positive significant effect for male subjects. Considering that men are on average more competitive than women, this result can be interpreted as support for the theoretical prediction that random grouping is beneficial when subjects have a competitive mindset and detrimental when subjects are non-competitive.

The knowledge of the effects of different reference points can be strategically used in firms and other institutions where relative performance feedback can be given. Our research shows that while the average reference point can evoke higher peer effects, the best reference point can create high incentives at the top of the distribution. The choice of the reference point thus depends on the objectives of the principal. The gender composition of the reference group should thereby be taken into account, since high reference points can demotivate women, but incentivize men to compete. Reference points might also differ with culture, as originally assumed in the theory paper by Thiemann (2017). Hence, people might rather compare to the top performers in competitive cultures, while people from non-competitive cultures might prefer the average performance as reference (Hofstede et al., 2010). In the experiment we are unable to select or group subjects according to their culture and are restricted to exogenously providing a reference point. Still, a principal should take into account the cultural background of the members of a group, when deciding about grouping policies or relative performance feedback. As our research shows the reaction to the reference point *does* depend on culturally differing factors like loss aversion or the tendency to select into competitive environments.

Both in the average and in the best treatment subjects positively react to the reference

point as shown by peer effects regressions. There is also evidence of non-linear peer effects with increasing sensitivity below the reference point and decreasing sensitivity above the reference point. The existence of these non-linear peer effects should logically induce effects of a regime change in the grouping procedure due to the change of the reference point. That this was not evident in this experiment might have several reasons: First, we have shown that a change of the grouping procedure has opposing effects on women and men. These effects might cancel out on the aggregate. Second, subjects might have been too aware of the existence of the other track. Participants knew that they were sorted into the low or high track, so that they could infer their relative position within the whole subject pool, rendering the exogenous change of the grouping procedure ineffectual. Third, as we have shown there is an immediate motivating effect of being sorted into the low track independent of the distance to the reference point (see also Kuhnen and Tymula, 2012; Gill et al., 2015). Thus, the tracking event itself might have outweighed the negative effect from the lower reference point.

In further research of grouping policies one might try to involve and disentangle the two channels through which the composition of groups might influence individual performance: mutual learning and norm setting. Our experiment only investigated the second channel by ruling out possibilities of learning from group members since subjects were isolated in cabins. Furthermore, our research has shown little effect of ability grouping, which might be due to subjects having been provided with too little information on their reference group. In a follow-up experiment subjects could be given information on the full group ranking and/or on the socio-economic and gender composition of the group. Further, incentives evoked by promotion/relegation from one track to the other might be analyzed.

Appendix

4.A Instructions

4.A.1 Original Version (German)

Herzlich Willkommen zum heutigen Experiment!
Sie nehmen heute an einem ökonomischen Experiment teil. Bitte beachten Sie, dass ab nun und während des gesamten Experiments keine Kommunikation gestattet ist. Wenn Sie während des Experiments irgendwelche Fragen haben, strecken Sie bitte Ihre Hand aus der Kabine. Einer der Experimentatoren kommt dann zu Ihnen. In diesem Experiment können Sie Geld verdienen, indem Sie Multiplikationsaufgaben lösen. Zur Lösung dieser Aufgaben dürfen Sie keinerlei Hilfsmittel verwenden, insbesondere kein Papier, Stift, Taschenrechner, Handy etc. Sollten Sie irgendein Hilfsmittel verwenden, werden Sie unmittelbar vom Experiment ausgeschlossen und erhalten keine Bezahlung. Dieses Experiment besteht insgesamt aus fünf Multiplikationsrunden von jeweils vier Minuten (240 Sekunden) Länge. Wir bitten sie in einer Runde so viele Multiplikationsaufgaben wie möglich zu lösen. Die Aufgaben bestehen immer aus der Multiplikation einer einstelligen Zahl mit einer zweistelligen Zahl. Sie bekommen eine Aufgabe so lange angezeigt bis Sie sie richtig beantwortet haben. Die Ihnen verbleibende Zeit wird im oberen Bildschirmrand in Sekunden angezeigt. Am Ende des Experiments wird zufällig eine der fünf Runden für die Auszahlung ausgewählt. Die Anzahl der in dieser Runde korrekt gelösten Aufgaben wird Ihnen entsprechend der folgenden Tauschrate ausgezahlt:

1 gelöste Aufgabe = 30 Cent

Zusätzlich bekommt jeder 5 Euro für die Teilnahme ausbezahlt. Zu Beginn des Experiments haben Sie die Möglichkeit in einer Proberunde von 30 Sekunden Länge die Bedienung des Eingabescreens zu testen. Nach Ablauf der fünf Runden bitten wir Sie noch einen kurzen Fragebogen zu beantworten. Das Experiment ist in drei Teile unterteilt. Teil 1 besteht aus einer wie oben beschriebenen Multiplikationsrunde.

[The order of the following two paragraphs was changed depending on the treatment]

Teil 2 [3] besteht aus Runde 2 und 3 [4 und 5]. Hier werden zu Beginn jeder Runde zufällig aus allen Teilnehmern Gruppen von fünf gebildet. Ihre Identität wird Ihren Mit-

spielern zu keinem Zeitpunkt bekannt gegeben. Vor jeder Runde erfahren Sie, wie hoch die durchschnittliche [beste] Leistung (in gelösten Aufgaben) unter Ihren Gruppenmitgliedern in der letzten Runde war.

Teil 3 [2] besteht aus Runde 4 und 5 [2 und 3]. Vor Runde 4 [2] werden Sie basierend auf Ihrer Leistung in Teil 1, entweder in Zug 1 oder Zug 2 eingeteilt. In Zug 1 befindet sich die Hälfte der Teilnehmer, die eine bessere Leistung als die Median-Leistung erbracht haben. In Zug 2 befindet die Hälfte der Teilnehmer, die eine schlechtere Leistung erbracht haben. Innerhalb dieser Züge werden wieder vor jeder Runde zufällig Gruppen mit fünf Mitgliedern gebildet. Zu Beginn von Teil 3 [2] erfahren Sie, in welchen Zug Sie eingeteilt wurden. Außerdem erfahren Sie wieder vor jeder Runde wie hoch die durchschnittliche [beste] Leistung unter Ihren Gruppenmitgliedern in der letzten Runde war.

Wenn Sie Fragen zu diesen Instruktionen haben, strecken Sie bitte jetzt Ihre Hand aus der Kabine. Einer der Experimentatoren kommt dann zu Ihnen.

Viel Erfolg!

4.A.2 English Translation

Welcome to today's experiment!
Today you are taking part in an economic experiment. Please note, that from now on and during the whole experiment no communication is allowed. If you have any questions during the experiment, please raise your hand and one of the experimenters will come to your cabin. In this experiment you can earn money by solving multiplication tasks. To solve the tasks you are not allowed to use any helping device, in particular no paper, pencil, calculator or mobile telephone. If you use any such helping device, you will be immediately excluded from the experiment and will get no remuneration. This experiment consists of five multiplication periods of four minutes each (240 seconds). We ask you to solve as many multiplication tasks as possible in one period. The tasks always consist of the multiplication of a one-digit number and a two-digit number. A task will be displayed as long as you need to answer the task correctly. Your remaining time will be displayed at the top of the screen. At the end of the experiment one of the five periods will be randomly chosen for the remuneration. The number of correctly answered problems in that period will be converted into Euros according to the following exchange rate:

1 solved problem = 30 Eurocent

In addition everyone receives 5 Euros for attendance. At the beginning of the experiment you will have the possibility to test the input-screen in a 30 seconds trial period. After

going through the five multiplication periods, we ask you to fill in a short questionnaire. The experiment is divided into three parts. Part 1 consists of one of the above described multiplication periods.

[The order of the following two paragraphs was changed depending on the treatment]

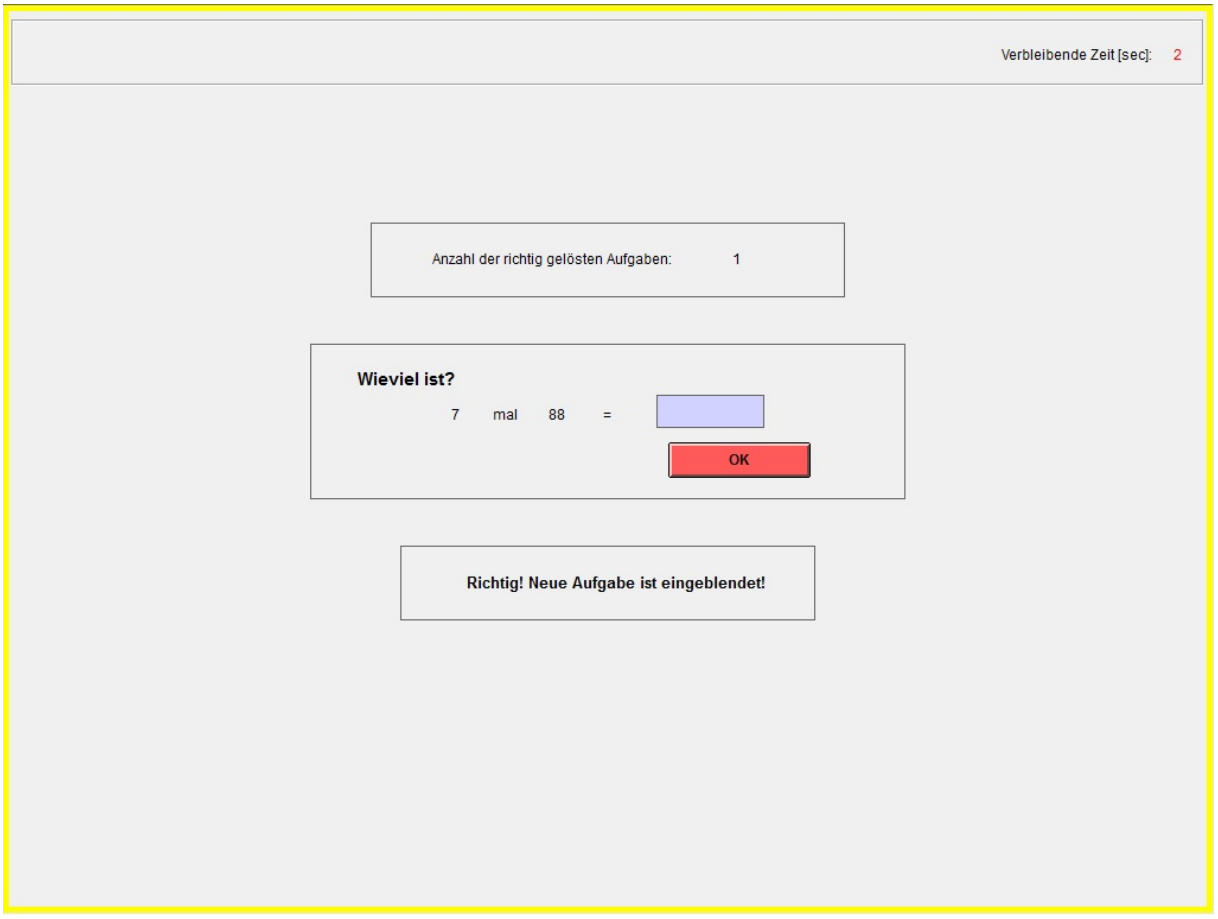
Part 2 [3] consists of periods 2 and 3 [4 and 5]. Here, you will be randomly allocated to a group of five. Your identity will at no point be published to your group members. Before each period you will receive information about the average [best] performance (in correctly answered problems) of your group members in the last period.

Part 3 [2] consists of periods 4 and 5 [2 and 3]. Before period 4 [2] you will be sorted either into track 1 or track 2 based on your performance in part 1. All the participants that performed higher than the median performance in the first period are allocated to track 1. Every subject that performed below median performance is allocated to track 2. Within these tracks again groups of five will be formed randomly before each period. At the beginning of part 3 [2] you will be told into which track you have been sorted. In addition you will again be informed before each period about the average [best] performance of your group members.

If you have questions about these instructions, please raise your hand out of your cabin. One of the experimenters will come to you.

Good luck!

4.B Input Screen Effort Task



4.C Questionnaire

1. How old are you? _____
2. What is your sex? □ Male □ Female
3. What are you studying? _____
4. What was your last math grade at your last school? □ 1 □ 2 □ 3 □ 4 □ 5 □ 6
5. When did you graduate from secondary school? _____
6. How much money do you have at your disposal per month? (including rent) □up to 400 Euro □ 400-600 Euro □ 600-800 Euro □ 800-1000 Euro □ 1000-1200 Euro □ more than 1200 Euro

7. Is German your native language? □ Yes □ No

8. If no, please indicate your native language? _____

9. Do you have the feeling that you could answer the multiplication problems faster over time due to practice? □ Yes, very much □ Yes, a little □ No

10. Did you get exhausted as time in the experiment went by, so that you could concentrate less? □ Yes, very much □ Yes, a little □ No

11. Imagine you are playing a quiz with 10 questions. Which possibility of earning money would you prefer? A: You get 4 Euro for each correct answer. B: You get 60 Euro , if you give more correct answers than another unknown person. How do you decide? □ A □ B

4.C.1 Loss Aversion

Loss aversion of subjects was assessed by a method developed by Abdellaoui et al. (2008). Subjects were asked the following three questions subsequent to the experiment:

1. Imagine a fair coin is flipped. You are offered a lottery, in which you can win 100 Euro if Head appears and nothing if Tails appears. Instead of playing the lottery you can accept a certain gain. Which of the following gains would you accept?

	reject	accept
10 Euro	□	□
20 Euro	□	□
30 Euro	□	□
40 Euro	□	□
50 Euro	□	□
60 Euro	□	□
70 Euro	□	□
80 Euro	□	□
90 Euro	□	□
100 Euro	□	□

2. The coin is flipped again. You are offered a game in which you lose 150 Euro if Head appears and lose 50 Euro if Tails appears. Alternatively you can accept a certain loss. Which of the following certain losses would you accept?

	reject	accept
-140 Euro	□	□
-130 Euro	□	□
-120 Euro	□	□
-110 Euro	□	□
-100 Euro	□	□
-90 Euro	□	□
-80 Euro	□	□
-70 Euro	□	□
-60 Euro	□	□
-50 Euro	□	□

3. The coin is flipped again. You can either reject the game and earn/lose nothing, or you can accept the proposed game. Which of the following games would you accept?

	reject	accept
If Head appears, you earn 30 Euro. If Tails appears you lose 50 Euro.	□	□
If Head appears, you earn 30 Euro. If Tails appears you lose 45 Euro.	□	□
If Head appears, you earn 30 Euro. If Tails appears you lose 40 Euro.	□	□
If Head appears, you earn 30 Euro. If Tails appears you lose 35 Euro.	□	□
If Head appears, you earn 30 Euro. If Tails appears you lose 30 Euro.	□	□
If Head appears, you earn 30 Euro. If Tails appears you lose 25 Euro.	□	□
If Head appears, you earn 30 Euro. If Tails appears you lose 20 Euro.	□	□
If Head appears, you earn 30 Euro. If Tails appears you lose 15 Euro.	□	□
If Head appears, you earn 30 Euro. If Tails appears you lose 10 Euro.	□	□
If Head appears, you earn 30 Euro. If Tails appears you lose 5 Euro.	□	□

The first question is used to elicit the participants' utility in the domain of gains. By presenting a gain prospect x_i its certainty equivalent G_i is elicited. From $u(G_i) = \delta^+ u(x_i)$ the δ^+ can be determined. The second question is used to elicit the certainty equivalent for losses L_i for a prospect of losses (x_i, y_i). With $u(L_i) = \delta^-(u(x_i) - u(y_i)) + u(y_i)$ the δ^- is determined. The third question serves the elicitation of an indifference loss L^* for a given gain G^*. Then the coefficient of loss aversion λ was determined from the following equation: $\delta^+ u(G^*) + \lambda \delta^- u(L^*) = u(0) = 0$. Throughout the elicitation linear utility functions were assumed. For a more detailed description of the procedure see Abdellaoui et al. (2008).

4.D Descriptive Statistics

Table 4.6: Summary Statistics

	Average			Best		
Variable	Mean	Std. Dev.	N	Mean	Std. Dev.	N
Number of Correct Answers	21.725	10.174	240	22.675	12.104	240
Refpoint	20.3	6.752	240	31.775	11.28	240
Loss Aversion	3.134	3.895	176	2.88	2.998	168
Competitiveness	0.283	0.452	240	0.267	0.443	240
Female	0.4	0.491	240	0.567	0.496	240
Math Grade	2.5	1.248	240	2.633	1.319	240
Years since Abitur	5.883	3.768	240	6.644	5.519	236
Studies Math	0.593	0.492	236	0.431	0.496	232
Age	24.817	3.796	240	25.7	6.402	240
Income	2.683	1.568	240	2.583	1.231	240
German Native Speaker	0.833	0.373	240	0.667	0.472	240

Table 4.7: Pairwise Correlations

Variable	NumberAns	Refpoint	LossAv.	Compet.	Female	Grade	Abitur
Number Ans.	1.000						
Refpoint	0.296***	1.000					
Loss Aversion	0.095**	0.038	1.000				
Competitiveness	0.079*	0.067	-0.172***	1.000			
Female	-0.128***	0.011	0.005	-0.073*	1.000		
Math Grade	-0.297***	-0.090**	0.116**	0.004	-0.128***	1.000	
Years since Abitur	-0.049	0.039	-0.072	-0.046	-0.210***	0.178***	1.000
Studies Math	0.187***	-0.069	0.081*	0.099**	-0.145***	0.009	-0.048
Age	-0.082**	0.005	-0.068	-0.076*	-0.215***	0.222***	0.932***
Income	0.055	-0.027	-0.060	0.147***	0.051	-0.019	0.174***
German Native	-0.050	-0.104**	-0.130***	-0.162***	-0.058	0.135***	0.131***

Notes: Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 4.8: Pairwise Correlations continued

Variable	StudMath	Age	Income	GermanNat.
Studies Math	1.000			
Age	-0.059	1.000		
Income	-0.024	0.151***	1.000	
German Native	-0.181***	0.116***	0.191***	1.000

Notes: Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Chapter 5

Culture as a Determinant of Intergenerational Education Mobility - Evidence from PISA

Abstract

This paper analyzes the determinants of cross-country differences in the impact of family background on student achievement. The focus among potential drivers is on country-specific family culture that can influence student motivation. Measures for culture are derived from questions in the World Values Survey about the valuation of hard work, competitiveness and the belief in free choice in life. In the first part of this paper we focus on native students to compare intergenerational mobility among more than 40 countries. In a second part data from students with immigration background is used in order to overcome endogeneity problems of the cultural variables. We find that disadvantages caused by family background can be overcome more easily in cultures with high beliefs in free choice. Especially male students also benefit if they come from competitive cultures. A high valuation of hard work, however, can decrease mobility.

JEL-Code: I20, I24, O15, J62

Keywords: student achievement, family background effect, inequality of opportunity, intergenerational education mobility, culture, PISA, World Values Survey, cross-country analysis, second generation immigrants

5.1 Introduction

Among the estimated determinants of student performance in school, the effect of family background (henceforth FB) is the largest. In international large scale assessments such as PISA (Programme for International Student Assessment) up to 30%[1] of the variance in student achievement can be explained by variables such as family income, parents' education or parents' occupation. This relationship is partly natural since parents pass on their genes to their children, resulting in a high correlation of innate ability between parents and children. Still, large cross-country differences in FB effects are evidence for other factors that drive the intergenerational mobility. These drivers are highly policy relevant, since relatively high estimated coefficients on FB variables are indicators for low equality of opportunity in a country, with the job-market success of children after school being pre-determined by the socio-economic background of their parents.

In the literature three main areas of potential drivers of cross-country differences in equality of opportunity have been identified (Solon, 2002; Corak, 2013): First, differences in labor market incentives, in particular the return to education, lead to differences in the parental investment in education. Second, public policy can influence the strength of the relationship through targeted distributive policies such as spending on early childhood education, health care or direct transfers to reduce disadvantages of children from low educated parents. Third, the influence of the family varies across countries. This influence is on the one hand through genes, where the strength of this mechanical heritability might be stronger in countries that practice intense assortative mating, i.e. marrying and reproduction within the same social class. On the other hand there are differences in family culture of how children are motivated to learn, what beliefs they have, which goals they thrive to achieve and in the means and skills of learning. The latter area has not been formally analyzed in the literature, in particular there are no theoretical models that describe the mechanisms, nor are there any approaches to estimate the impact of cultural differences. This paper aims at identifying some concrete cultural characteristics that impact intergenerational mobility and at quantifying their influence.

In the first part of the paper, we use extensive data from the PISA 2012 study to estimate the influence of FB, proxied by the average number of years of parents' education, on the achievement of their children for 64 countries. The coefficient on parents' education is a measure for the intergenerational education mobility, with a high coefficient indicating low mobility since student achievement is predetermined by their parents' education. We

[1] Own estimation with PISA 2012 data, using as FB variables parents' years of education, parents' wealth, books at home and HISEI, an index of the highest occupational status of the parents.

further interact parents' education with certain country level variables to explain the cross country differences in FB effects. Among these are variables on national culture of competitiveness, the valuation of hard work and the belief in free choice and control, all derived from questions within the World Values Survey (WVS). In the second part of the paper, we address concerns about the endogeneity of these cultural variables, due to reverse causality between mobility and culture, by looking at immigrants only. Since an immigrant's culture stems from their origin country, it is exogenous to the level of mobility in the destination country. Here we merge PISA data from the 2006, 2009 and 2012 waves to gather enough observations from second generation immigrants. Regressing the achievement of these immigrant students on their parents' education in interaction with the aforementioned cultural variables, yields insight on the causal effect of national culture on student mobility.

We find that the impact of FB on student achievement in highest in Oceania, Russia and Latin America and lowest in Scandinavia, Southeast Asia and Central Europe[2]. The analysis of native students suggests that the impact of FB is lower in countries where people have competitive mindsets and believe in free choice and control over their life. The analysis of second generation immigrants can establish causality of the mobility increasing effect of the belief in free choice. For competitiveness this can only be confirmed for male students. Surprisingly, for immigrants from countries where hard work is valued the effect of FB on student achievement is higher.

This paper is organized as follows. Section 5.2 gives an overview of the related literature. In Section 5.3 the achievement data, the proxy for FB and the cultural variables are described. Section 5.4 presents the estimation strategy and the results of the estimation with native students from PISA 2012. Section 5.5 presents estimation strategy and results of the estimation with second generation immigrants. In Section 5.6 some robustness checks are given. Section 5.7 concludes.

5.2 Related Literature

This paper relates to two main fields of economic research. First, the topic of FB effects relates to the economic literature on intergenerational earnings or education mobility and equality of opportunity. Second, our immigrant approach in the second part of the paper relates to research using the epidemiological approach.

The first field of research we relate to is the extensive literature on intergenerational earnings mobility, which analyzes the relationship of earnings between parents and their

[2]The countries summarized under these regions can be taken from Table 5.9 in Appendix 5.C.

children. Typically researchers estimate the elasticity of intergenerational mobility by running a regression of a sons earnings on his fathers earnings (see e.g. Black and Devereux, 2011; Corak, 2006). A society where this elasticity is close to one is interpreted to be highly immobile with children from disadvantaged backgrounds having little opportunity of social advancement. Closer related to this paper is research that investigates intergenerational education mobility, looking at the relationship of the years of education or qualifications of parents and their children (e.g. Nimubona and Vencatachellum, 2007; Aydemir et al., 2013). Because of the strong relationship between education and earnings, there should be a strong mapping between the two measures of mobility. Our research even goes one step back in looking at achievement data of children still at school and their parents' years of education. The advantage lies in the extensive and internationally comparable data available from large scale assessments such as PISA. In addition, the data is very up-to-date compared to data on earnings that can only be measured once the children have entered the labor market.

Within this field papers of particular interest to us, are those that study the cross-country differences in intergenerational mobility. Hertz et al. (2007) study the trends in the intergenerational transmission of education for a sample of 42 countries. They find that while the regression coefficient of parents' education has markedly decreased over the past 50 years, the correlation between parents' and children's education has not. As a reason for this they identify the increased variance of attained years of education. Chevalier et al. (2003) compare different measures of educational mobility for 20 countries, finding an inverse relationship between mobility and equality. Solon (2002) summarizes international findings of intergenerational earnings mobility and develops a theoretical framework for interpreting cross-country differences. Consistently cross-country research shows that Scandinavian countries are highly mobile, while low mobility is found in Latin American countries. Central Europe has medium levels and the UK and the USA rather low levels of mobility.

The potential drivers of these cross country differences have also been studied. In his theory consisting of utility-maximizing parents who invest in their child's human capital, Solon (2002) finds that the intergenerational earnings elasticity depends positively on both the "strength of the mechanical heritability of income-generating traits" (p.65) and the earnings return to human capital investment. Furthermore, it varies inversely with the progressivity of government investment in children's human capital. Corak (2013) wrote a survey on different impact factors on differences in earnings mobility. He also identifies the labor market, public policy and investment in human capital within the family as potential factors. Also Blanden (2013) authored a survey on different impact

factors, focusing on the relationship between mobility and the Gini coefficient, the return to education and public spending on education. Finally, Ichino et al. (2011) sketches a theory of the parameters that determine the intergenerational elasticity of income, taking into account private and collective decisions. He points out that public expenditure on education is an endogenous impact factor, since it is determined by societal preferences for redistribution, societal heterogeneity and the strength of cultural transmission. He underlines his theory by calculating correlations of the elasticity with proxies for the mentioned exogenous factors, namely the political participation of low income vs. high income voters, ethnolinguistic fragmentation and an index of weak family ties.

All the aforementioned studies calculate correlations of potential determinants with the elasticity of earnings or education mobility measured by country. Schütz, Ursprung, and Wößmann (2008) go a step further in estimating education production functions on an individual student level with TIMSS data, taking as indicator for FB the number of books at the home of the students. They estimate the impact of institutional factors such as pre-school enrollment, age of first tracking and educational expenditure per student and can thereby control for some other factors that might impact mobility and are correlated with FB. Our contribution to this literature is twofold. First, building on the work by Ichino et al. (2011), we believe that intergenerational mobility is driven to a high degree by societal and cultural preferences. We investigate the relationship of some of these cultural characteristics with FB effects on student achievement. Second, instead of only calculating correlations we follow the approach by Schütz, Ursprung, and Wößmann (2008) in estimating education production functions. And, in the second part of the paper, we can even rule out endogeneity of cultural variables with respect to intergenerational mobility by using the epidemiological approach.

The epidemiological approach is a cure for the problem of small data and endogeneity problems when analyzing cross country differences. This approach uses data from immigrants whose culture is exogenous to characteristics of the country they live in. A seminal paper was written by Carroll et al. (1994) who argue that country differences in saving rates can be explained by cultural differences, underlining this point with the saving behavior of immigrants in Canada that differs according to origin country. Luttmer and Singhal (2011) show that culture is an important determinant of preferences for redistribution using survey data of immigrants in 32 destination countries. Closely related to our research within this field is a study by Isphording et al. (2015), who also use data from immigrants within the PISA waves 2003, 2006, 2009, and 2012 to show that there is a strong causal effect of reading performance on math performance.

5.3 Data

5.3.1 Achievement

Data on student achievement is taken from the PISA study, where math, science and reading knowledge of 15-year-old students from OECD and partner countries is tested in three year intervals. The individual student assessment is measured on a scale that is based on a mean for OECD countries of 500 points and a standard deviation of 100 points that were set in PISA 2003 when the first PISA scale was developed (OECD, 2013b). The sampling procedure of PISA is a two-stage sampling design, where in each country, first, a non-random sample of schools is selected, then a simple random sample of 35 students from the 15-year-old student population from each school is drawn (OECD, 2014). We use PISA data since it provides a huge database that is comparable across countries, which is a big advantage to other studies on intergenerational mobility that merge survey data from different countries (e.g. Hertz et al., 2007). In particular studies analyzing intergenerational *earnings* mobility find it hard to get adequate and comparable data across countries (Corak, 2006, p.6). We use achievement data from *mathematics* since it is generally viewed as being most comparable across countries (Hanushek, Link, and Wößmann, 2013).

For the purposes of the first part of this paper we use data from the most recent PISA wave in 2012. For the regional comparisons of mobility we are able to include all 64 countries from PISA 2012, but some countries are dropped in later analysis that includes cultural variables due to lacking data. The proxy for FB that we use is most likely correlated with the immigration status of students. In addition, immigrants have a different cultural background than native students, such that we cannot ascribe national culture to them. Consequently, we drop all students from the sample who themselves were not born in the country of the test, or whose parents were not born in the country of the test. The mean test score in mathematics across all countries in our sample is 453 score-points with a standard deviation of 104. Table 5.9 in Appendix 5.C lists all countries and the number of student observations used in the first part of this paper as well as the mean test score of the country.

For the second part of the paper we look at second generation immigrants only. We focus on second instead of first generation immigrants to avoid selection problems due to motives for emigration. In this analysis a second generation immigrant is a student who was born in the test country, but whose parents were born in a different country. We exclude those whose parents were born in two different countries, since culture could then

not be correctly assigned, assuming that culture is transmitted from parents to children. The PISA survey asks students where they themselves, their mother and their father were born, but many students answered with "in a different country" or "in another European country". All these observations are dropped. In order to gather enough observations the PISA 2012 data is pooled with the PISA 2009 and PISA 2006 data, with roughly a third from the generated sample from each wave. Furthermore, all destination country and origin country pairs with less than 10 student observations were dropped in order to get valid sample sizes for each country pair. This leaves us with roughly 20.000 students from 29 destination countries and 39 origin countries, resulting in 85 country pairs (see Appendix 5.D for observation numbers by destination and origin country). The mean test score among the second generation immigrants in our sample is 481 score-points with a standard deviation of 105.

5.3.2 Family Background

As an indicator for students' FB we take average parents' education in years. The reasons are, first, to ensure comparability to other studies of intergenerational education mobility that use similar measures. Instead of using only the father's years of education (e.g. Checchi et al., 2013), we use the average of the mother and father (similarly in Hertz et al., 2007; Chevalier et al., 2003), since the mother usually plays a crucial role in the upbringing and education of children. Second, other measures like "books at home" as used in Schütz et al. (2008) might explain more variance in the achievement data, but they vary between countries for reasons not correlated to FB, e.g. cultures' differing appreciation of books as status symbols. Third, compared to using income as a FB proxy, parents' education has a more direct effect on children's achievement in terms of role modeling and quality time at home. In contrast to studies of intergenerational *earnings* mobility, there is also no problem of women working less because of assortative mating and labor supply reasons (Black and Devereux, 2011, p.17).

In the PISA survey students were asked for the educational qualification of their father and their mother. Using information on the regular duration of the different qualification levels in each country from ISCED (International Standard Classification of Education) (OECD, 2013b, p.260), we can derive the years of education from father and mother, ignoring any possible repetitions at school or university and calculate the simple average. In the case where students only report one parental qualification, e.g. because they are raised by a single parent, we use this information only without calculating an average.

Figure 5.1 shows box plots of the generated variable across regions from the PISA 2012

sample.[3] The variance in average parents' education within countries is on high levels in every country (the minimum standard deviation is 1.2 in Russia and Kazakhstan, the maximum is 4.8 in Tunisia). There are concerns that in some (developing) countries many parents might have no education at all, resulting in a coarseness in the data (Hertz et al., 2007). However, in our sample there is no country with more than 10% of parents with an average education of 0. The PISA data is rather new, which is an advantage in terms of women's education, i.e. we find no regular differences in the distribution of fathers and mothers education.

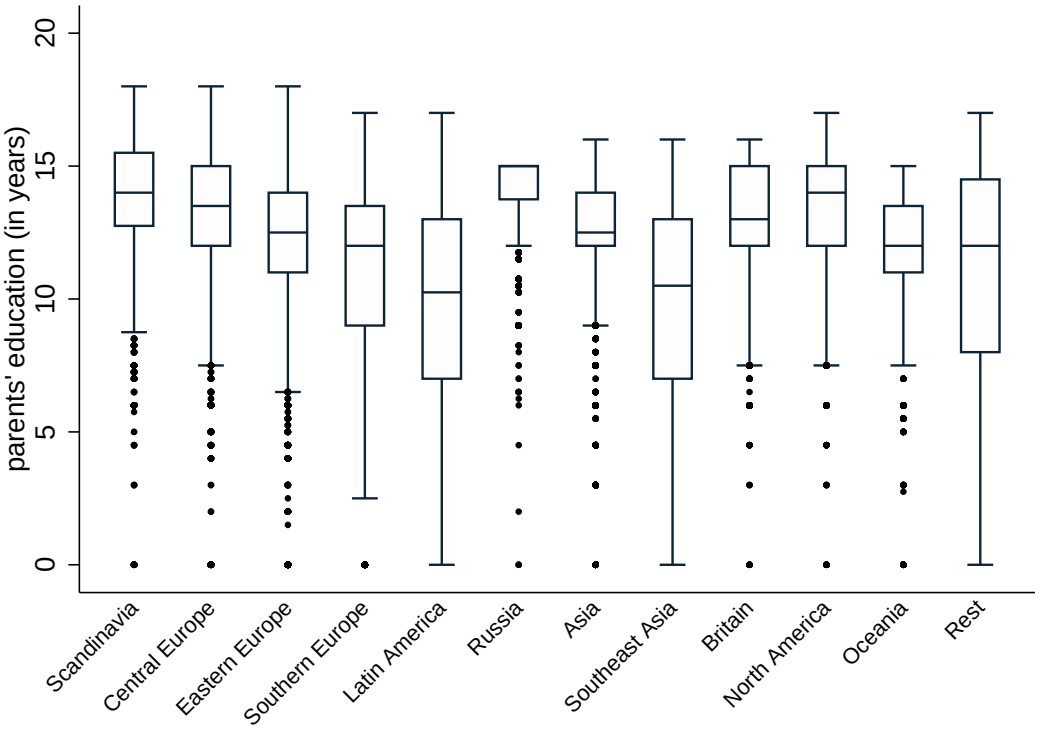


Figure 5.1: Box Plots of Average Parents' Education by Region, Native Students

When looking at second generation immigrants only, parents' qualifications have been converted into years of education according to the ISCED conversion rate of the origin country (OECD, 2010, p. 178; OECD, 2013b, p. 260). Figure 5.2 shows the distribution of average parents' education across destination countries. We see e.g. that Australia only allows immigration of well-educated people, while Germany and the Netherlands allow for immigrants with diverse background.

In the estimation we do not control for any other variables that capture aspects of FB, since we want to to capture the whole effect of FB by our variable of average parents'

[3]The box plot reports the median, the 2nd and 3rd quartiles and the whiskers are found at the last observation within the 1.5 × interquartile range.

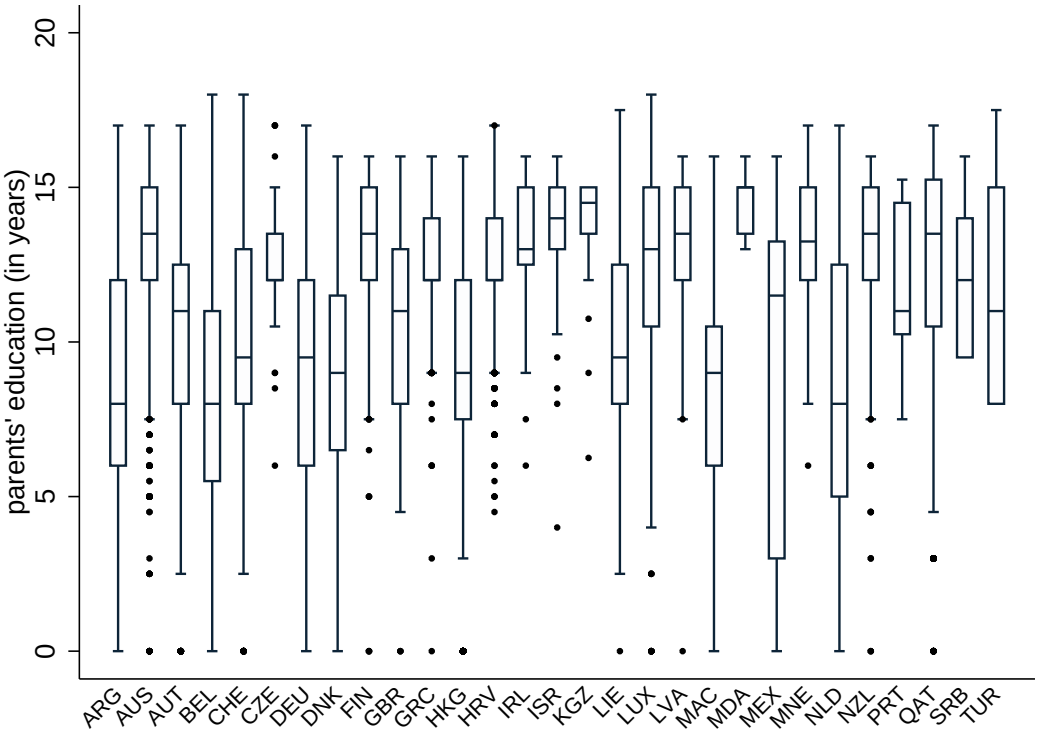


Figure 5.2: Box Plots of Average Parents' Education by Destination Country, Second Generation Immigrants

education. Parental education can influence a child's achievement through many channels. Most importantly, parents with high education levels transmit their genes to their children, who are thus more likely to have higher innate ability. Second, parents with high education will probably have higher incomes, with which they can afford better schools and other enrichment investments for their children. Third, educated parents serve as role models and may spend quality time with their children in which these are educated also outside school. Other factors that come with educated parents might be better neighborhoods that lead to positive peer effects and better connections to schools and possible employers (Corak, 2013). If these mechanisms work with the same strength in every country, they cannot be among the causes for cross-country differences in mobility.

5.3.3 Culture

In their equilibrium theory of the distribution of income and intergenerational mobility Becker and Tomes (1979, p.1158) assume that children receive endowments of capital from their parents that among others include the "learning, skills, goals, and other 'family commodities' acquired through belonging to a particular family culture". Corak (2013, p.90) describes non-monetary investments of families in their children that reflect "the

development of behavior, motivation, and aspirations". There are no more detailed descriptions or formal analyzes of cultural endowments in the literature that could provide grounds to derive hypothesis for our purposes. Based on these scarce descriptions and the availability of data, we chose three variables of culture from the World Values Survey (Inglehart, 2014) that we believe are likely to cause cross-country differences in educational mobility: Opinions on hard work as an important child quality, the appraisal of competition and the belief in free choice in life. We take data from the 4th, 5th and 6th WVS wave, which were conducted between 1999 and 2014. We calculate per country and wave averages of the answers of all interviewees to the questions of interest, of which we take the latest available data point to merge with the PISA data. These country level indicators for culture are assumed to reflect time invariant cultural values of the native people in the countries that are transmitted from parents to children. To confirm the time persistence of the cultural values, we run panel data regressions of the cultural variables that vary across country and time on country and time dummies. Performing F-tests on the time dummies proves that these are insignificant (p-values above 0.1) for all three cultural variables.

The variable on "hard work" is derived from a question that focuses on the transmission of values from parents to their children: *"Here is a list of qualities that children can be encouraged to learn at home. Which, if any, do you consider to be especially important? Please choose up to five: Good Manners, Independence, Hard Work, Feeling of Responsibility, Imagination, Tolerance and respect for other people, Thrift saving money and things, Determination and perseverance, Religious faith, Unselfishness, Obedience."* The deduced country level variable *hardwork* can be interpreted as the fraction of survey participants that mentioned the particular quality as important. We believe a higher valuation of hard work in a society might reduce FB effects, since hard working students from disadvantaged backgrounds might catch up the advance that students from better background have. A positive effect of the absolute value of *hardwork* on mobility would thus mean that this value has a higher return for students from low educated parents than for students with high educated parents. In addition we computed a ratio, i.e. the share of low educated participants (people with education up to finished secondary school) that mentioned the quality divided by the share of highly educated participants (people with some university education or a university degree), such that the ratio is bigger than one if lower educated people place more value on hard work than higher educated people. We expect the ratio to be negatively correlated with FB effects, since children from lower educated families might catch up with those of higher educated background if they place more value on hard work. Figure 5.3 shows the shares of high and lower educated survey

participants that mentioned hard work as an important quality across regions. In every region lower educated people place more value on hard work than the higher educated, but differences are generally bigger between than within countries.

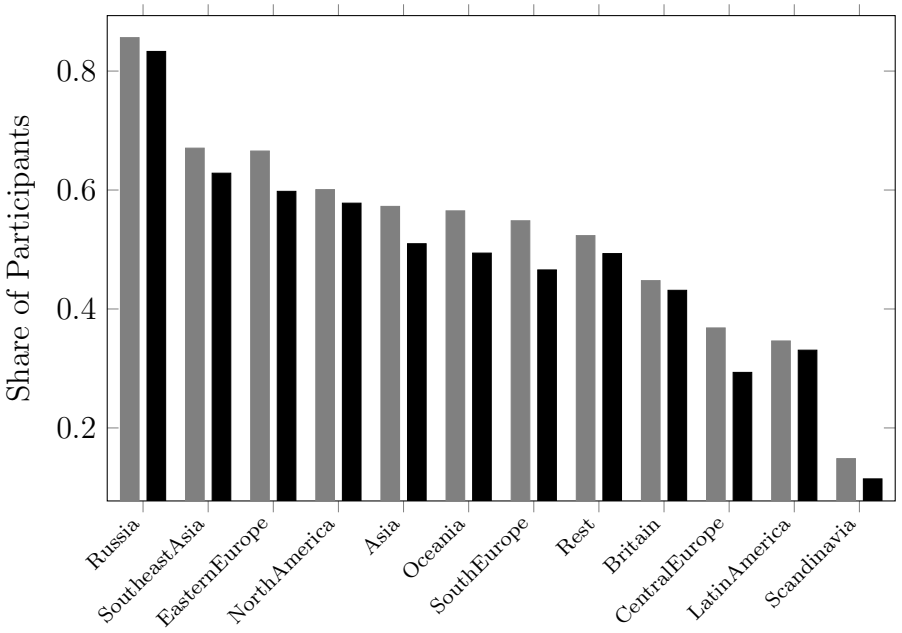


Figure 5.3: Share of Participants that Mentioned Hard Work as Important Child Quality among Participants with Low Education (Gray) and with High Education (Black)

Furthermore, we believe that competitiveness, i.e. the desire to socially compare and outperform your peers in class, can help to overcome disadvantages from FB. To assess values on competitiveness, answers to the following statement were taken: "*Competition is good. It stimulates people to work hard and develop new ideas" vs. "Competition is harmful. It brings the worst in people*". People were asked to place their view about this statement on a scale from 1 to 10, where 1 means "*competition is good*" and 10 means "*competition is harmful*". As a country level variable the average answer per country is calculated, normalized and reverse coded so that 1 is the most competitive country and 0 the least competitive in the sample. We believe that the absolute value of competitiveness is negatively correlated with FB effects, since competitiveness induces positive peer effects. For instance if also students from disadvantaged backgrounds have the desire to compare their performance to good students, their motivation is positively influenced. This can lead to positive dynamics (for a theory on this see Thiemann, 2017). We again also compute the ratio of competitiveness between low educated and high educated people. In countries where low educated people are more competitive than high educated people, intergenerational mobility might be higher due to students from disadvantaged backgrounds

gaining higher motivation from comparing to well-performing peers. Figure 5.4 shows that mostly better educated people are more competitive.

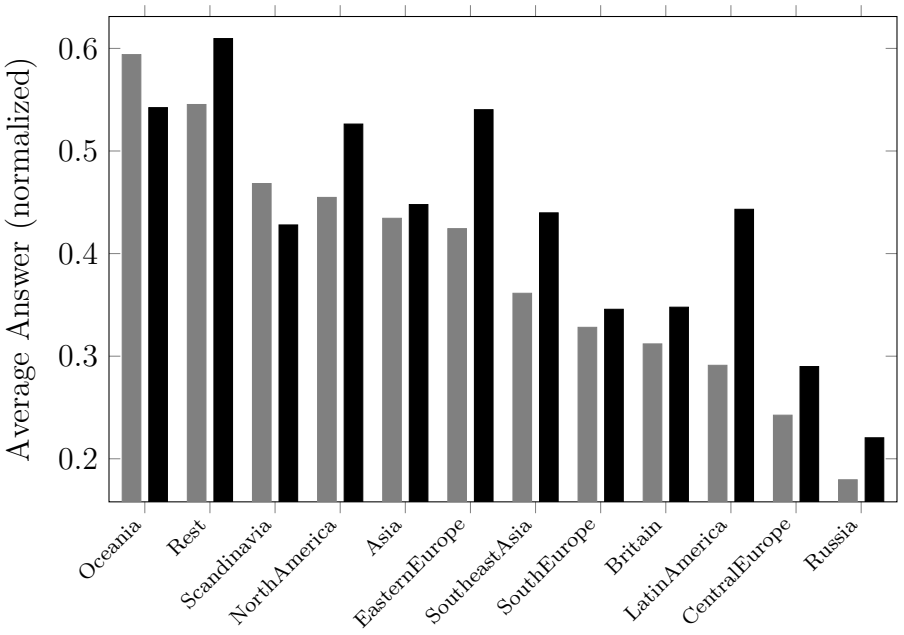


Figure 5.4: Average Answer to the Question on Competitiveness of Participants with Low Education (Gray) and with High Education (Black)

Third, a child's motivation and aspirations depend on their beliefs about how they can influence their future life by being successful at school. We thus take a question from the WVS on people's beliefs on how much control and free choice they think they have over their lives: *"Some people feel they have completely free choice and control over their lives, while other people feel that what they do has no real effect on what happens to them. Please use this scale where 1 means "none at all" and 10 means "a great deal" to indicate how much freedom of choice and control you feel you have over the way your life turns out."* The aggregate country level variable is the average answer, which is again normalized to take on values between 0 (no belief in free choice) and 1 (full belief in free choice). We expect a negative relationship of the generated variable *freechoice* with FB effects, since students who believe that they have full control over their future career, have a higher motivation and can thus more easily overcome disadvantages from FB. Typically we can think of people who believe in the American Dream, i.e. that they can make it from "rags to riches", in contrast to the caste-system in South Asia where the caste that you are born into determines your future profession and social life. A significant positive effect on mobility of the absolute value of this variable would again indicate that the belief in free choice has a higher return for students with low educated parents. We also compute

the ratio as for the other two cultural variables to capture the effect of differences in the motivation between students from high educated parents and low educated parents. From Figure 5.5 you can see that for instance in Australia and New Zealand (Oceania) low educated people believe more in free choice than high educated people, which should lead to a higher motivation of disadvantaged students and potentially higher intergenerational mobility.

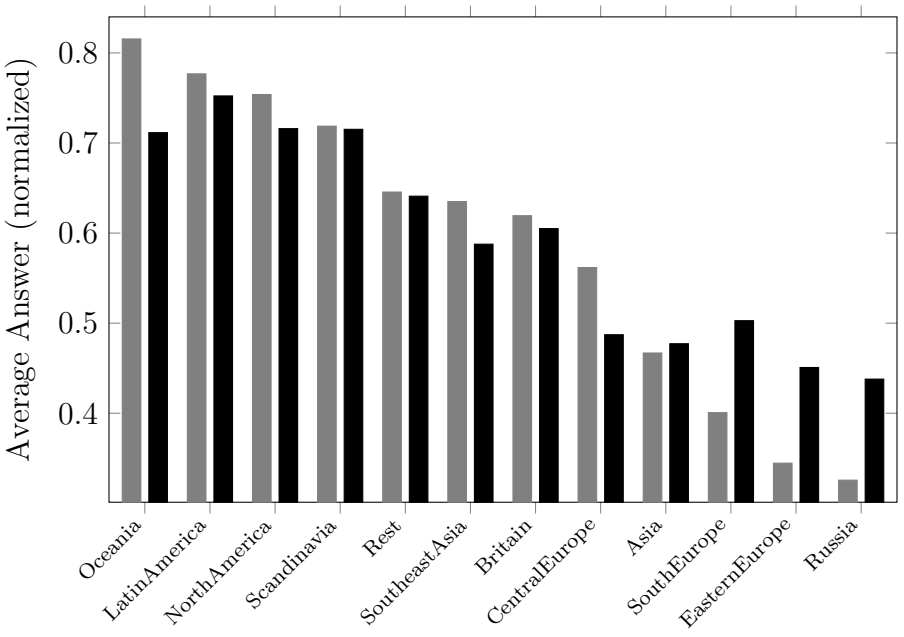


Figure 5.5: Average Answer to the Question on Free Choice in Life of Participants with Low Education (Gray) and with High Education (Black)

The PISA data is also merged with other country level data on the development status, education policy and religious adherence and pairwise correlations are calculated to explore what the derived cultural variables measure (data sources are found in Appendix 5.A and correlations in Appendix 5.B). Accordingly, *hardwork* is negatively correlated with per capita GDP (-0.383***) and the share of Protestants (-0.517***) and Catholics (-0.410***) in the country. This reflects that people from poorer countries place more importance on values of survival that come with hard work, i.e. economic and physical security, rather than on values of self-expression (see e.g. Inglehart, 1997). The *ratio of hardwork* is negatively correlated with income inequality (measured by the GINI index) in the country (-0.304**), indicating that the bigger income inequality the less do lower educated people emphasize hard work. *Competitiveness* is negatively correlated with the share of Catholics (-0.325**) and positively with the share of Muslims in a country (0.261*). The *ratio of competitiveness* has a high positive and significant correlation with

the PISA mean test score (0.453***) and per capita GDP (0.371***), indicating that in wealthier countries lower educated people do believe more in the beneficial effect of competition than higher educated people. *Freechoice* is positively correlated with the share of Protestants (0.329**) as well as with the GINI index (0.301**), which shows that people from more unequal countries in terms of income do believe more in free choice. The *ratio of freechoice* is highly correlated with the absolute value of *freechoice* (0.655***), suggesting that countries where low educated people believe more in free choice than high educated people are also the countries with high absolute values of *freechoice*. The ratio is also positively correlated with per capita GDP (0.398***). Figure 5.6 shows correlations of the derived variables *hardwork*, *competitiveness* and *freechoice* with the the GDP per capita in 2012, of which only the correlation with *hardwork* is significant (-0.383***).

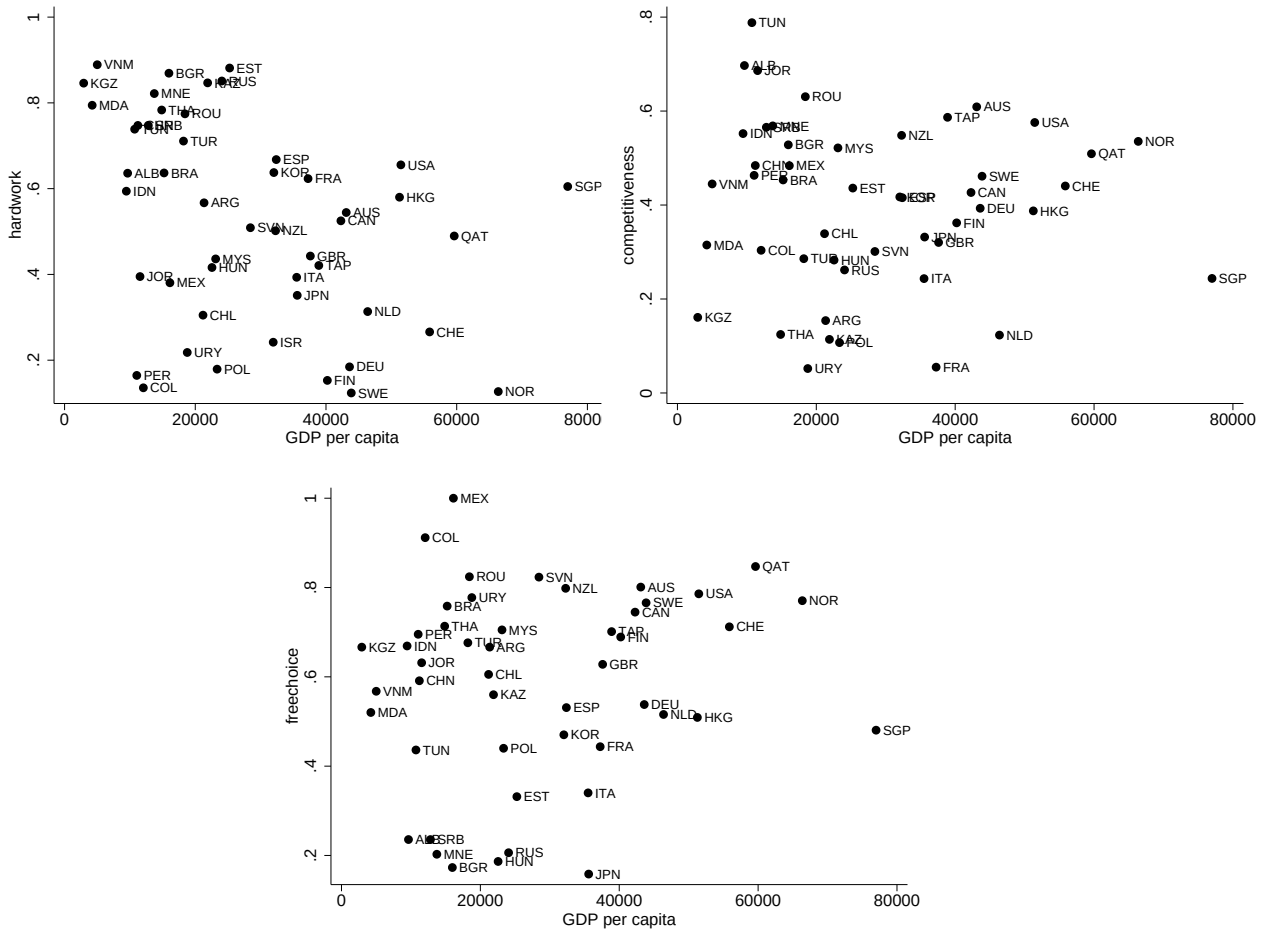


Figure 5.6: Correlations of Cultural Variables with GDP per Capita 2012 (World Bank, 2015a)

5.4 Part I: Native Students

5.4.1 Estimation Strategy

The basis for the estimations in this first part is a pooled Ordinary Least Squares (OLS) model as given in Equation (5.1). A_{ic} is math achievement of student i as achieved in the PISA test 2012. The main independent variable of interest is $pared_{ic}$, which is the average number of years of education of the student's parents. The vector $\mathbf{C}_c$ includes country level variables that may explain cross-country differences in FB effects, such as public policy, labor market and our variables of culture. Interactions of these variables with $pared_{ic}$ are also included, so that the coefficients δ indicate how much the country level variable increases or decreases the effect of parents' education on student achievement. The vector $\mathbf{S}_{ic}$ includes student level control variables.

$$A_{ic} = \alpha + \beta pared_{ic} + \mathbf{C}_c\gamma + (\mathbf{C}_c \times pared_{ic})\delta + \mathbf{S}_{ic}\kappa + \epsilon_{ic} \tag{5.1}$$

In a second step country fixed effects μ_c are included to account for all country-specific factors that might influence student achievement in PISA. The vector $\mathbf{C}_c$ drops from the regression due to perfect multicollinearity with the fixed effects, but $\mathbf{C}_c$ can stay in the interaction. For the identification of Equation (5.2) the assumption of no unobserved cross-country heterogeneity can be replaced by the less restrictive assumption that there is no unobserved cross-country heterogeneity that is related to the size of the FB effects.

$$A_{ic} = \alpha + \beta pared_{ic} + (\mathbf{C}_c \times pared_{ic})\delta + \mathbf{S}_{ic}\kappa + \mu_c + \epsilon_{ic} \tag{5.2}$$

Parents' education is considered to be exogenous with respect to student achievement, since it is usually determined before the children go to school. For the identification of β, i.e. the effect of parents' education on achievement, we could control for many factors that are correlated with parents' education to avoid omitted variable bias, such as parents' occupation, better neighborhoods or parents' income. However, we want to capture all these effects in the coefficient on $pared_{ic}$ as to get a general indicator of the full impact of FB. If all the aforementioned factors are correlated with parents' education in every country to the same degree, β as a measure for the full FB effect will not be biased. For parents' income some tests that we perform show that the correlation of parental income and parental education is higher in countries with lower GDP per capita. However, in

later analysis it is controlled for this by including GDP per capita and the return to education as control variables within the country level vector $\mathbf{C}_c$.[4]

Another omitted variable that is probably highly correlated with parents' education is the students' innate ability. Since innate ability is a result of the genes that have been passed on by parents, it can be interpreted as a part of FB. The correlation of innate ability and parents' education might differ between countries, since more developed countries that make schooling available for all layers of the population probably have a higher correlation between the two. We cannot control for innate ability, but in order to identify δ we control for the rate of out-of-school children in primary education in 1985 within the vector $\mathbf{C}_c$ to mirror the available schooling that the parents of our 15-year-old students had. We believe that countries where the out-of-school rate was low seized more of the ability potential of their population such that in 2012 the correlation between student innate ability and parents' education is higher.

The estimation intentionally does not control for school characteristics, such as schools resources, endowments or institutional features, as in standard education production functions. Again, this is because we are interested in the total impact of FB including any effect that might work through parents' differential access to schools, their school choice or their influence on school policies. The vector $\mathbf{S}_{ic}$ thus only includes students' age, gender and a dummy on whether they live with a single parent. These are exogenous variables that might influence student achievement, but should not be correlated with parents' education. They could be dropped, but we include them in order to be consistent with the existing literature (Nimubona and Vencatachellum, 2007; Schütz et al., 2008). As parents' education is not an average of both parents, when the student is growing up with a single parent, we also control for this state. The correlations of all control variables with parents' education and achievement are given in Appendix 5.C.

For the identification of δ, i.e. the effect of country level variables on intergenerational mobility, it is important to control for factors that might determine country differences in mobility and are systematically correlated with the cultural variables of interest in C_c. We discuss the country level control variables that we use in more detail in Section 5.4.2.2. There might still be problems of endogeneity of some cultural variables with respect to mobility. For instance some might argue that the belief in free choice depends to a huge degree on whether there *is* mobility between the generations in a country. Also the existence of high levels of mobility might lead to people working harder. Since, however, the dependent variable is the *individual* achievement score of 15-year-old students, rather

[4] As a robustness check we also perform the analysis from part I including "family wealth" as a student level control variable. The results are very similar to those presented here with only small differences in the size of the coefficients.

than a mobility measure of a country, the endogeneity problem should be minimal. Also, these students do not choose the country they are born in, which renders self-selection problems non-existent. To find cure for the remaining endogeneity problems, we look at immigrants only in the second part of the paper. Their mobility level in the destination country can then not be associated with the origin country variable of culture.

The underlying data is hierarchical data, where students are nested in schools and schools are nested in countries. Since our primary interest is in the identification of an interaction between a student level and a country level variable, we identify the countries as our primary sampling unit. Therefore standard errors are clustered at the country level, such that standard errors are measured as if there were only as many observations in the regression as there are countries. Furthermore, student sampling weights provided by PISA are adapted such that they give equal weight to each country in addition to the weighting of students' sampling probability within each country.

5.4.2 Results

5.4.2.1 Regions

Table 5.1 aims at giving an overview of how intergenerational education mobility varies between regions. Therefore we performed regressions as in Equation (5.2) with vector C_c only consisting of dummies for the regions the countries belong to. The reference region in the estimation is Scandinavia. From the literature on intergenerational education mobility we expect Scandinavian countries to have high degrees of mobility, whereas Latin American and Anglo-Saxon countries are expected to have low degrees of mobility. Central European countries are expected to have medium levels (Corak, 2013; Hertz et al., 2007; Blanden, 2013). In a second regression we also control for the PISA mean test score of each country. This is because the results for the coefficients on parents' education for countries scoring high in the PISA test are generally inflated because of level effects.

In specification (1) of Table 5.1 a baseline estimation including only parents' education and control variables is given. Across countries the average impact of parents' education on student achievement is given by the coefficient 6.8, i.e. if average parents' education increases by one year, students on average perform 6.8 score-points higher. This can roughly be interpreted as 6.8% of the standard deviation of achievement. Put differently, a student with parents who have no education is on average one standard deviation (100 score-points) worse than a student with parents who have a university degree (15 years of education).

Table 5.1: Effect of Parents' Education on Student Achievement in Regions

Variables	(1)	(2)	(3)
Parents' Education (in years)	6.810***	7.143***	-16.595***
	(0.105)	(0.395)	(1.405)
Parents' Educ. × Asia		1.682***	-1.456**
		(0.575)	(0.604)
Parents' Educ. × Latin America		-0.950**	3.922***
		(0.440)	(0.530)
Parents' Educ. × Eastern Europe		0.889*	3.132***
		(0.503)	(0.508)
Parents' Educ. × Russia		2.609**	4.518***
		(1.276)	(1.300)
Parents' Educ. × Southeast Asia		-1.354***	0.690
		(0.503)	(0.515)
Parents' Educ. × Central Europe		0.352	-0.392
		(0.681)	(0.680)
Parents' Educ. × North America		1.468**	1.146*
		(0.626)	(0.626)
Parents' Educ. × Rest		-2.361***	1.874***
		(0.516)	(0.586)
Parents' Educ. × Oceania		5.685***	5.462***
		(0.788)	(0.787)
Parents' Educ. × Britain		1.019	1.011
		(0.837)	(0.838)
Parents' Educ. × Southern Europe		0.274	1.129**
		(0.506)	(0.502)
Parents' Educ. × Mean Test Score (in 100 points)			4.754***
			(0.271)
Student controls	Yes	Yes	Yes
Country FE	Yes	Yes	Yes
Avrg. R^2	0.36	0.36	0.36
Country obs.	64	64	64
Student obs.	373,315	373,315	373,315

Notes: Dependent Variable: Math achievement in PISA 2012. Included control variables: age, female and single parent. Pooled OLS regression adjusted for student weights. Standard errors are clustered at the country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

From specification (2) and (3) we see that the inclusion of the PISA mean test score changes the coefficients on the interactions of average parents' education and the regions to a high degree. Regions with high PISA scores have a much lower mobility, suggesting that controlling for the mean test score is necessary due to inflated coefficients. In specification (3) only Asia has a higher mobility than Scandinavia with a 1.5 points lower average achievement increase per student if parents have one more year in education. Oceania, Russia and Latin America have the lowest levels of intergenerational mobility. In Australia and New Zealand (Oceania), for instance, the performance increase for students whose

parents have one more year in education is 5.5 score-points higher than for students in Scandinavia. Medium levels of mobility are found in North America, Eastern and Southern Europe. In Central Europe and Southeast Asia no significant difference to Scandinavia is found.

The rather low estimates for Latin American and Southeast Asian countries can also be explained by a small selection bias. Especially in these countries we find rather high rates of out-of-school children in secondary education of up to 15% (UNESCO Institute for Statistics, 2016b). Especially these children are probably those with little opportunity of climbing up the social ladder.

5.4.2.2 Culture

In a next step we include interactions of parents' education with our culture variables. Table 5.2 reports the results in specification (1). In the following specifications we include other country level variables in interactions with parents' education to control for factors that might also impact country differences in FB effects and are correlated with culture. In column (2) we include school policy variables, namely the fraction of children that receive pre-primary schooling, the squared of this variable, public spending on primary education and a dummy on whether the country has a tracked school system. According to Schütz et al. (2008) we expect pre-primary schooling to have an inverse u-shape relation with the effect of FB. Tracking is supposed to decrease mobility, since students from poor backgrounds have a higher probability of being in tracks with less talented peer groups (Pekkarinen et al., 2009). Public spending on primary education should decrease the effect of FB since it translates into investment in the human capital of children from all backgrounds (Solon, 2002). In specification (3) we include the returns to education[5] as a control for labor market incentives. These are expected to increase the effect of parental years of education, since high income parents invest more in their children, if payoffs to education are higher (Solon, 2002). Additionally, it is commonly assumed that the returns are higher for high-ability students (Black and Devereux, 2011). Due to the lack of observations in this variable, it is reported in a separate specification. In column (4) we add controls for macro level variables, namely per capita GDP and the Gini index. These are supposed to impact the effect of FB not in a direct way, but they are correlated because of policy variables and labor market circumstances that come with these. Income inequality, measured by the Gini index, is supposed to have a positive correlation with FB effects, because it is harder to overcome disadvantages in very unequal societies (Corak,

[5]The return to education is measured as a coefficient on years of education in a regression of earnings on years of education per country by Psacharopoulos and Patrinos (2004).

Table 5.2: Culture and Intergenerational Education Mobility

Variables	(1)	(2)	(3)	(4)
Parents' Educ.	21.966***	26.740***	14.968***	17.891***
	(3.844)	(4.841)	(4.247)	(5.543)
Parents' Educ. × Hardwork	-1.811***	-1.083	-3.434***	-1.835**
	(0.591)	(0.661)	(0.603)	(0.730)
Parents' Educ. × Ratio of Hardwork	-0.706	-2.601***	2.865***	-1.384*
	(0.682)	(0.852)	(0.683)	(0.768)
Parents' Educ. × Competitiveness	-5.094***	-4.098***	-0.901	-3.806***
	(0.586)	(0.745)	(0.881)	(0.745)
Parents' Educ. × Ratio of Comp.	1.080	-9.797**	-0.454	-4.839
	(3.239)	(4.570)	(3.875)	(4.823)
Parents' Educ. × Freechoice	0.462	-0.927	-5.673***	-1.269
	(0.692)	(0.852)	(0.915)	(0.829)
Parents' Educ. × Ratio of Freechoice	-26.765***	-18.706***	-16.658***	-19.482***
	(4.539)	(5.039)	(6.147)	(5.090)
Parents' Educ. × Tracking		0.143		0.328***
		(0.116)		(0.127)
Parents' Educ. × Pre-Primary Enrollment (%)		0.051		0.012
		(0.035)		(0.038)
Parents' Educ. × Pre-primary squared		-0.000		-0.000
		(0.000)		(0.000)
Parents' Educ. × Exp. on Prim. Educ. (% of cGDP)		-0.064***		-0.038*
		(0.020)		(0.021)
Parents' Educ. × Out-of-School Children 1985 (%)		-0.090***		-0.114***
		(0.024)		(0.026)
Parents' Educ. × Return to Education			0.276***	
			(0.054)	
Parents' Educ. × GDP per capita (in 1000 USD)				-0.006
				(0.008)
Parents' Educ. × Gini Index				0.095***
				(0.023)
Parents' Educ. × Mean Test Score (in 100 points)	2.895***	2.846***	1.915***	3.146***
	(0.185)	(0.287)	(0.231)	(0.303)
Student Controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Avrg. R^2	0.40	0.41	0.38	0.40
Country obs.	48	41	35	38
Student obs.	310,630	274,223	262,927	267,485

Notes: Dependent Variable: Math achievement in PISA 2012. Included control variables: age, female and single parent. Pooled OLS regression adjusted for student weights. Standard errors are adjusted for clustering at the country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

2013). That countries with higher GDP have per se higher FB effects can be explained by the higher innate ability potential of children from lower socioeconomic backgrounds in developing and emerging countries. Again we control for level effects by including

the interaction of parents' education with the PISA mean test score of the country. We do not control for religious adherence, since religions are considered as the roots and providers of various morals and values, but we want to single out the particular cultural value itself that drives the difference across countries. For detailed data sources refer to Appendix 5.A.

Most of the control variables have the expected signs, except for pre-primary enrollment, which has a non-significant effect. Ability tracking is associated with lower mobility and public spending on education with higher mobility. Since we control for the rate of out-of-school children the coefficient on GDP per capita is insignificant (see specification (4)). Overall, the results yield strong evidence of the importance of cultural aspects in intergenerational education mobility. First, there is strong evidence that students in competitive countries overcome their disadvantages from FB more easily. In the country with the highest *competitiveness* in our sample, Yemen, the effect of 1 year more in parents' education is estimated to be almost 3.8 score-points lower (most restricted model in specification (4)) than in South Africa, the country with the lowest *competitiveness*. There is no evidence that the *ratio of competitiveness* matters for FB effects. Second, the ratio of *freechoice* is more important in reducing FB effects than its absolute value. The ratio is highest in Switzerland (1.006), where lower educated people believe more in free choice than higher educated people. In Hungary, the country with the lowest ratio (0.837), higher educated people believe more in free choice. Our model predicts that in a country where this variable is 0.1 higher, the effect of an increase of parents' education by one year on student achievement is 2 score-points lower (from specification (4)). Third, the variable *hardwork* is mostly associated with lower FB effects. In specification (4) we find evidence that mobility is higher in countries where more people think that *hardwork* is an important child quality. The ratio of this variable yields inconsistent results.

5.5 Part II: Second Generation Immigrants

In order to find casual evidence on the importance of family culture, this Section focuses on second generation immigrants. Analyzing the FB effect among immigrants yields the advantage of the students' cultural background, which stems from their origin country, being exogenous to the level of education mobility in the destination country. Furthermore, this approach increases the variance in the cultural variables since it is possible to assign a different culture to each immigrant, rendering our primary unit of observation country-pairs instead of countries. Additionally, we can analyze whether the distance of immigrant culture to the destination country culture has an impact on how students can

offset disadvantages from FB. This might be important, since the culture of the destination country influences institutions, like teaching styles or grading systems in school, and peer effects from native students.

5.5.1 Estimation Strategy

Identification of the influence of origin country culture and cultural differences is assessed with a pooled OLS model as in Equation (5.3) and 5.4 respectively.

$$A_{ijkt} = \alpha + \beta pared_{ijkt} + (pared_{ijkt} \times \mathbf{CUL}_j)\delta + \mathbf{S}_{ijkt}\kappa + \mu_t + \mu_{jk} + \epsilon_{ijkt} \tag{5.3}$$

$$A_{ijkt} = \alpha + \beta pared_{ijkt} + (pared_{ijkt} \times \mathbf{CULDIST}_{jk})\delta + \mathbf{S}_{ijkt}\kappa + \mu_t + \mu_{jk} + \epsilon_{ijkt} \tag{5.4}$$

The dependent variable A_{ijkt} is the PISA math score of student i from country j that was tested in country k in year t. The student characteristics vector $\mathbf{S}_{ijkt}$ consists of student's age, gender and singleparent. The regression as in Equation (5.3) includes a vector of the student's origins country culture $\mathbf{CUL}_j$ in an interaction with parents' education. Because of perfect collinearity with the country pair fixed effects μ_{jk}, the vector $\mathbf{CUL}_j$ does not exist its own in the regression. The vector $\mathbf{CUL}_j$ consists of the variables *hardwork*, *competitiveness* and *freechoice*. Compared to Part I, we do not include the ratios of the cultural variables, since these can explain differences in FB effects on an aggregate country level, but not differences on an individual migrant level.

In the second model as in Equation (5.4) we include the distance between the origin culture of the immigrant and the culture of the test country $\mathbf{CULDIST}_{jk} = \mathbf{CUL}_j - \mathbf{CUL}_k$ instead of just the student's origins country culture. The distance is, for instance, positive, if the immigrant comes from a country where hard work is more appreciated than in the destination country. In both specifications our interest is in the coefficients δ. In Equation (5.3) they indicate the change in the influence of one more year in parents' education on student achievement if the student comes from a country that scores 1 compared to 0 in the particular cultural variable. In Equation (5.4) they indicate the change in the impact if the student comes from a country that scores 1 point higher in the particular cultural variable than the destination country.

We include dummies for the year of the test μ_t in order to control for unobserved year fixed effects or differences in the test design between 2006, 2009 and 2012. Fixed effects for the country pairs μ_{jk} are included, rendering origin or destination country dummies unnecessary. Note that the country pair fixed effects differ according to the direction of migration, i.e. $\mu_{jk} \neq \mu_{kj}$. They thus also control for any origin or destination

country characteristic that might be systematically correlated with parents' education. For example high levels of unemployment in the area of unqualified work might be a reason for emigration resulting in a selection of rather low educated emigrants. Also the potential problem of a systematic selection of immigrants depending on the country of origin (e.g. some countries might be in war) is accounted for by the fixed effects.

For the identification of β in both models there might be factors that are correlated with parents' education to different degrees between the countries of origin. In terms of income, for instance, some parents with high education might have only little income if their origin country diplomas are not recognized. Parents' education might thus be a biased estimator of the full FB effect. PISA data does not provide a variable for parents' income, but as a proxy it offers an index of family wealth, which is based on students' responses on whether they have particular items at home, e.g. the number of computers, cars or rooms with a bath or shower. We include this variable as a control in all following estimations.

For the identification of δ in Equation (5.3) and (5.4) we need to assume that there are no omitted origin country or country pair characteristics that are correlated with the cultural variables and could influence the mobility of immigrant students in order to avoid omitted variable bias. One country pair characteristic that could increase immigrants mobility through faster assimilation is whether the countries share a common language. We control for this in specification (3) and (4) in Table 5.3. An origin country characteristic that could influence migrants' educational mobility in the destination country is how the role of women is understood in the culture of the origin country. If women are not expected to work, but to fulfill traditional household duties, investments in the education of girls are probably low and likewise the educational level of the mother. We thus control for female employment as a percentage of the female population 15+ in the origin country in specification (4) to proxy for this aspect. Another origin country characteristic we control for in this specification is the rate of out-of-school children in 1985 in primary education in the origin country. We think that this could impact intergenerational mobility of the migrants, since immigrant children from countries with lower out-of-school rates probably have a lower potential due to innate ability than children from countries with higher rates.

Some might argue that immigrants in general have a higher motivation to work hard and believe more in free choice and are thus not a random sample of the distribution of beliefs and preferences in the country of origin. If this, however, is equally true for immigrants from all origin countries it should not bias our results. In addition the selection problem is the main reason for conducting the analysis with second generation immigrants, since these are not affected by the original reasons for emigration.

We follow Isphording et al. (2015) in adjusting the standard errors for clusters at the country pair level. This helps to avoid biases due to skewness in the number of observations (for instance there are more than one thousand immigrants from China to Hong Kong, while there are only 11 from China to the United Kingdom).

5.5.2 Results

Table 5.3 presents the estimates of Equation (5.3), i.e. the effect of immigrant culture on their educational mobility. The estimates of Equation (5.4), i.e. the effect of cultural distance on educational mobility among immigrants, are presented in Table 5.4.

From specification (1) in Table 5.3 we see that among second generation immigrants one year more in parents' education increases student achievement in the PISA test by 3.3 score-points. Immigrant students are thus much more mobile than native students (see Section 5.4.2.1), which is in line with the results from Aydemir et al. (2013), who study immigrants' educational mobility in Canada. In specification (3) we see that once we control for common language, the significance of the cultural variables increases (for

Table 5.3: Second Generation Immigrants

Variables	(1)	(2)	(3)	(4)
Parents' Educ.	3.287***	6.025**	3.160	2.024
	(0.592)	(2.716)	(2.718)	(3.243)
Parents' Educ. × Hardwork_j		4.707*	9.929***	8.918**
		(2.802)	(3.167)	(3.575)
Parents' Educ. × Comp_j		-4.157	-5.018*	-2.109
		(2.818)	(2.833)	(3.048)
Parents' Educ. × Freechoice_j		-7.381*	-9.477**	-9.627**
		(3.904)	(4.013)	(3.753)
Parents' Educ. × Comlang			5.146***	5.281***
			(1.436)	(1.635)
Parents' Educ. × Female Employment_j				0.039
				(0.048)
Parents' Educ. × Out-of-school Children85_j				-0.134**
				(0.065)
Student Controls	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Country Pair FE	Yes	Yes	Yes	Yes
Avrg. R^2	0.39	0.39	0.39	0.40
Test Country obs.	29	29	29	29
Origin Country obs.	39	39	39	38
Student obs.	20,272	20,272	20,272	19,953

Notes: Dependent Variable: Math achievement in PISA 2006, 2009 and 2012. Included control variables: age, female, singleparent and wealth. Pooled OLS regression adjusted for student weights. Standard errors are clustered at the test country × origin country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

correlations of control variables see Appendix 5.D). Common language is highly significant, indicating that the FB effect is much higher for immigrants from countries with the same language. As expected immigrants from countries with higher rates of out-of-school children have lower FB effects.

From specification (3) and (4) in Table 5.3 we see that immigrants from countries where hard work is valued more have lower levels of mobility. From Table 5.4 it is obvious that in particular the distance in the valuation of hard work between the origin and the destination country increases the impact of parents' education on achievement. For a student from a country, where everyone considers hard work to be an important child quality who is tested in a country where no-one does so, the increase in achievement of one more year in parents' education is on average 14 score-points higher compared to a student from a country where no-one considers hard work to be important (see specification (3)). A possible explanation is that the return to *hardwork* is higher for immigrant students from well-educated parents, such that this cultural value aggravates inequality of opportunity. This is a contrast to the results from part I, where the coefficient

Table 5.4: Cultural Distance

Variables	(1)	(2)	(3)
Parents' Educ.	1.801	-2.085	-5.582*
	(1.126)	(2.016)	(3.310)
Parents' Educ. × Dist. Hardwork	4.043	13.511***	13.991***
	(2.837)	(4.838)	(4.865)
Parents' Educ. × Dist. Comp	-4.813**	-5.044**	-2.300
	(2.367)	(2.408)	(3.401)
Parents' Educ. × Dist. Freechoice	-4.022	-5.117	-7.459**
	(3.444)	(3.511)	(3.255)
Parents' Educ. × Comlang		7.881***	8.595***
		(2.593)	(2.639)
Parents' Educ. × Female Employment_j			0.078
			(0.049)
Parents' Educ. × Out-of-school Children85_j (%)			-0.080
			(0.090)
Student Controls	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Country Pair FE	Yes	Yes	Yes
Avrg. R^2	0.46	0.46	0.47
Test Country obs.	16	16	16
Origin Country obs.	35	35	34
Student obs.	8,765	8,765	8,667

Notes: Dependent Variable: Math achievement in PISA 2006, 2009 and 2012. Included control variables: age, female, singleparent and wealth. Pooled OLS regression adjusted for student weights. Standard errors are clustered at the test country × origin country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

on *hardwork* was probably downward biased.

There is, furthermore, strong evidence that mobility is higher for immigrants from countries where people believe in free choice and control over their lives. This evidence becomes stronger once we control for common language. The cultural *distance in free-choice* is less important, since this value does not impact the social interaction with native students. From specification (4) in Table 5.3 we see that an increase of 1 year in parents' education among students that come from the country with the highest level of *freechoice*, Mexico, leads to an average increase in achievement by 10 score-points less than in the country with the lowest belief in free choice, India.

In terms of *competitiveness* there is some evidence that the *distance in competitiveness* between origin and test country can decrease the impact of parents' education. The importance of the distance is reasonable here, since competitiveness is a character trait that influences the interaction with native peers. Once we control for female employment the significance vanishes (see specification (3) in Table 5.4). Female employment is negatively correlated with *competitiveness*, indicating that the employment rate of women is smaller in more competitive countries. The mobility increasing effect from *competitiveness* in specification (2) of Table 5.4 might thus stem from young female students using their new opportunities in the destination country, when their mothers did not have this opportunity in the home country. We also conduct the estimation of Equation (5.3) and (5.4) only for male students (for results see Appendix 5.D.1). The evidence here is much stronger and the coefficients are bigger for all cultural variables. This is especially true for *competitiveness* that is highly significant also after controlling for the female employment rate and the out-of-school children rate.

5.6 Robustness Checks

To test robustness, we first do the same analysis as in part I and II with a different measure for FB, namely the amount of books that is found at the student's home. This measure is used by Schütz et al. (2008) who argue that the *books* measure is more readily comparable across countries, since the meaningfulness of educational qualifications varies to a high degree between countries. In addition to the included individual control variables in Equation (5.2), we also control for the number of people in the household, since this might be a reason for a higher number of books that is not related to the socio-economic background of the parents. Results for all estimations using *books* as a proxy for FB are given in Appendix 5.E. The negative impact of the *ratio of freechoice*, *competitiveness* and *hardwork* on FB effects among native students can be confirmed. Using books as a

proxy for FB in the estimation with second generation immigrants shows that the effect of *competitiveness* and the *distance of competitiveness* is even larger and robust to the inclusion of control variables. The same is true for the effect of the *distance in hardwork*. *Freechoice* is only significant after controlling for the rate of out-of-school children as a proxy for the higher ability potential of origin countries. *Books* might be a better proxy for the innate ability of parents, where parents' education is rather a measure for the socio-economic circumstances the parents live in. Disadvantages from socio-economic background can thus rather be overcome by the belief in free choice.

Another major robustness check that we conduct is a two step procedure with the data from part I. We first estimate per country coefficients on average parents' education only controlling for gender, age and singleparent. In a second step we regress the estimated coefficients on the same explanatory country level variables as in Table 5.2. The estimated coefficients on parents' education vary from 21.3 in Poland to 1.9 in Qatar, which can be interpreted as the average increase in score-points if average parents' education increases by one year. Poland is thus the least mobile country in our sample. Estimates in Albania and Liechtenstein are not significantly different from zero and are dropped from further analysis. Table 5.21 in Appendix 5.E, shows that the significant negative coefficient on *freechoice* and *competitiveness* can mostly be confirmed.

Another concern is the linearity of the effect of parents' education. Adding a squared version of average parents' education to the estimation as in Equation (5.2) leads to a significant, but very low positive coefficient on this variable (0.192^{***}) indicating that the quadratic effect is negligible.

We ran the analysis with second generation immigrants as in Equation (5.3) and (5.4) also with first generation immigrants, i.e. students who themselves and their parents were born in a different country than the test country. We excluded students who were more than 12 years old when they arrived in the country of the test, since the PISA results of these students rather mirrors the schooling in their home country. The results presented in Table 5.23 in Appendix 5.E confirm the results on *competitiveness* from Table 5.4. The mobility increasing effect of *freechoice*, especially the distance in the belief between test and origin country, can also be confirmed. *Hardwork*, however, is not significant in this regression, maybe because of a selection bias of first generation immigrants. If first generation immigrants from countries with a low value of *hardwork* are more hard-working than measured by this variable, then the coefficient estimated here is downward biased.

5.7 Conclusion

In this paper we analyze the influence of cultural variables on intergenerational education mobility. We interact country level variables on the valuation of hard work, the views on competition and the belief in free choice with parents' years of education to measure their common influence on student achievement in an education production function framework. In a first step we run estimations using only native students to be able to compare the influence of parents' education across countries and regions. In line with existing literature, we find a low impact of parents' education on student achievement in Asian countries (e.g. China, Japan, Korea) and Scandinavian countries, as well as a high impact in Russia, Latin America and Oceania. In terms of culture, we find that the influence of parents' education on achievement is lower in countries where the view of competition is more positive and the belief in free choice is higher. To find a causal relationship between these cultural variables and mobility, we repeat the analysis using only data from second generation immigrants. Their cultural background is exogenous to the mobility in the country of the test. Here the mobility increasing effect of the belief in free choice can be confirmed. *Competitiveness* only significantly increases mobility for male students. The valuation of hard work, however, can decrease mobility, probably because of higher returns to this quality for children of more educated parents.

To the best of our knowledge this paper is the first approach to measure the effect of national culture on the intergenerational mobility in education. Looking at the estimated coefficients the effects of culture are relatively big in size and potentially more important than policy variables, such as public expenditure on education or the tracking system in a country. Our results are of importance for policy makers, not only in education politics, but also in labor politics, because of the strong relationship between educational and income mobility. Since mobility and thus equality of opportunity depend on cultural aspects, the political focus should not only be on the design of adequate incentives in school and the labor market, but also on the formation of values and beliefs in early childhood. Children from disadvantaged backgrounds can be motivated by the belief that they have free choice and control over their lives and by a competitive environment. In particular, our research also shows that immigrant students lag behind native students (e.g. second generation immigrants in Germany lag behind native students by on average almost half a standard deviation[6]) also because of their cultural background. They can be motivated not by emphasizing hard work, but by fostering their beliefs that they are in control of their own destiny and have all opportunities.

[6]Own estimation with PISA 2012 data.

There is much more research to be done, especially in isolating and defining, as well as measuring cultural values that are important for the creation of a mobile society. Also, the exact mechanism behind the impact of certain cultural values on mobility could be described in a theoretical framework.

Appendix

5.A Data Description and Sources

Variable	Definition	Source
Tracking	Dummy= 1 if country has a tracked school system at the age of 15	OECD (2013a, p.78)
Pre-primary	Gross Enrollment Ratio in Pre-Primary Education, both sexes (in %) 2002	UNESCO Institute for Statistics (2016a)
Return to education	Return to investment in education (coefficient on years of education in a regression of earnings on years of education) for different years	Psacharopoulos and Patrinos (2004)
per capita GDP	GDP per capita (PPP, current international $) 2012	World Bank (2015a): World Development Indicators
Exp. on Prim. Educ.	Government expenditure per student, primary (% of GDP per capita)	World Bank (2015b): World Development Indicators
Mean Test Score	PISA 2012 mean country score in mathematics	OECD (2013b)
Gini Index	Gini Index (World Bank Estimate) (2012 or latest available up to 2007)	World Bank (2016a): World Development Indicators
Religious adherence	Shares of Protestants, Catholics, Muslims, Buddhists, Hindus and non-religious by country	Association of Religion Data Archives (ARDA) (2016): World Religion Dataset
Comlang	Dummy=1 if a countrypair has the same official language	Melitz and Toubal (2014)
Female_Employ	Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate) in 2012	World Bank (2016b) World Development Indicators
Out-of-School Children 1985	Rate of out-of-school children of primary school age, both sexes (%) in 1985	UNESCO Institute for Statistics (2016b)

5.B Culture Variables from WVS

Table 5.5: Country Values of Cultural Variables, generated from WVS

country	cnt	hardwork	hardworkratio	comp	compratio	freechoice	freechoiceratio
Albania	ALB	0.636	1.325	0.697	0.922	0.236	0.928
Argentina	ARG	0.567	0.872	0.154	1.001	0.667	0.990
Australia	AUS	0.544	1.162	0.609	1.003	0.801	0.992
Bulgaria	BGR	0.869	1.088	0.528	0.956	0.173	0.862
Bosnia	BIH	0.538	1.352	0.672	0.943	0.342	0.912
Belarus	BLR	0.877	0.988	0.381	0.961	0.299	0.987
Brazil	BRA	0.637	1.386	0.453	0.920	0.758	0.975
Canada	CAN	0.525	1.038	0.427	0.964	0.745	0.969
Switzerland	CHE	0.266	1.287	0.441	0.982	0.712	1.006
Chile	CHL	0.305	0.941	0.339	0.906	0.605	0.968
China	CHN	0.747	1.042	0.484	0.996	0.591	0.940
Colombia	COL	0.136	0.911	0.304	0.841	0.912	0.959
Cyprus	CYP	0.471	1.077	0.395	1.013	0.728	1.004
Germany	DEU	0.184	1.010	0.393	0.970	0.538	0.982
Ecuador	ECU	0.443	1.136	0.435	0.918	0.818	0.933
Egypt	EGY	0.439	0.924	0.831	0.949	0.326	0.888
Spain	ESP	0.668	1.158	0.416	0.989	0.531	0.946
Estonia	EST	0.881	1.036	0.436	0.931	0.332	0.896
Finland	FIN	0.153	1.215	0.362	0.997	0.689	0.956
France	FRA	0.623	1.378	0.055	0.975	0.444	1.000
Great Britain	GBR	0.443	1.038	0.320	0.973	0.628	0.961
Hong Kong	HKG	0.580	1.334	0.388	1.006	0.509	0.975
Hungary	HUN	0.416	1.000	0.283	0.876	0.187	0.837
Indonesia	IDN	0.594	1.062	0.552	0.924	0.669	0.968
India	IND	0.626	1.291	0.187	1.175	0.000	0.954
Iraq	IRQ	0.578	0.972	0.694	1.000	0.400	0.903
Italy	ITA	0.393	1.211	0.243	0.975	0.340	0.872
Jordan	JOR	0.395	1.123	0.686	0.913	0.632	0.942
Japan	JPN	0.351	0.784	0.332	0.936	0.158	0.905
Kazakhstan	KAZ	0.847	0.999	0.114	0.982	0.560	0.903
Kyrgyzstan	KGZ	0.846	0.978	0.161	0.932	0.666	0.949
South Korea	KOR	0.638	1.047	0.417	0.984	0.470	0.947
Lebanon	LBN	0.398	0.879	0.295	0.965	0.538	0.950
Morocco	MAR	0.671	1.106	0.745	1.020	0.290	0.874
Moldova	MDA	0.794	1.079	0.315	0.952	0.520	0.907
Mexico	MEX	0.381	0.985	0.484	0.889	1.000	0.955
Montenegro	MNE	0.822	1.049	0.568	0.952	0.203	0.904
Malaysia	MYS	0.436	0.987	0.522	0.904	0.705	0.984
Netherlands	NLD	0.313	1.176	0.123	0.950	0.516	0.963
Norway	NOR	0.127	1.187	0.535	0.996	0.771	0.961
New Zealand	NZL	0.502	1.124	0.548	1.004	0.798	1.002
Pakistan	PAK	0.557	0.978	0.247	0.909	0.641	0.943
Peru	PER	0.164	1.001	0.463	0.956	0.695	0.961
Philippines	PHL	0.693	1.118	0.342	0.986	0.649	0.955
Poland	POL	0.179	1.195	0.107	0.902	0.440	0.939
Palestine	PSE	0.430	0.987	0.654	0.997	0.504	0.941
Qatar	QAT	0.490	1.126	0.509	0.934	0.847	0.987
Romania	ROU	0.774	1.188	0.631	0.906	0.824	0.946

Table 5.6: (continued)

country	cnt	hardwork	hardworkratio	comp	compratio	freechoice	freechoiceratio
Russia	RUS	0.851	1.057	0.262	0.974	0.206	0.905
Singapore	SGP	0.605	1.095	0.244	0.963	0.481	0.999
Serbia	SRB	0.748	1.101	0.566	0.938	0.236	0.915
Slovenia	SVN	0.509	1.166	0.301	0.929	0.823	0.919
Sweden	SWE	0.124	1.546	0.461	1.026	0.766	0.951
Taiwan	TAP	0.421	1.601	0.587	0.969	0.701	0.976
Thailand	THA	0.783	1.091	0.125	0.966	0.713	0.972
Tunisia	TUN	0.739	1.056	0.788	0.979	0.436	0.929
Turkey	TUR	0.711	1.041	0.286	0.960	0.676	0.965
Ukraine	UKR	0.856	1.024	0.229	0.912	0.406	0.944
Uruguay	URY	0.218	1.226	0.052	0.860	0.777	0.947
United States	USA	0.655	1.040	0.576	0.928	0.786	0.971
Uzbekistan	UZB	0.927	1.002	0.682	0.930	0.797	0.955
Vietnam	VNM	0.889	1.073	0.445	0.985	0.568	0.954
Yemen	YEM	0.317	1.104	1.000	0.958	0.361	0.906
South Africa	ZAF	0.692	0.952	0.000	0.990	0.603	0.931

Table 5.7: Summary Statistics of Country Level Variables

Variable	Mean	Std. Dev.	Min.	Max.	N
Hardwork	0.525	0.237	0.124	0.889	49
Ratio of Hardwork	1.113	0.157	0.784	1.601	49
Competitiveness	0.398	0.178	0.052	0.788	48
Ratio of Comp.	0.952	0.041	0.841	1.026	48
Freechoice	0.584	0.214	0.158	1	48
Ratio of Freechoice	0.949	0.037	0.837	1.006	48
Tracking	0.255	0.441	0	1	47
Pre-Primary Enrollment	70.515	24.14	7.388	113.184	62
Exp. on Prim. Educ. (% of cGDP)	19.661	9.109	0.313	54.157	43
Out-of-school Children 1985 (%)	7.745	6.041	0.085	32.352	62
Return to Educ.	8.544	3.119	2.7	14.7	36
Mean Test Score (in 100 points)	4.661	0.643	3.31	6.13	49
GDP per capita (in 1000 USD)	28.113	17.069	2.921	76.988	49
Gini Index	35.888	7.325	25.59	53.54	45
Share of Protestants	0.127	0.194	0	0.806	46
Share of Catholics	0.263	0.31	0	0.877	46
Share of Muslims	0.181	0.305	0	0.99	46
Share of Buddhists	0.056	0.17	0	0.87	46
Share of Hindus	0.01	0.04	0	0.265	46
Share of no Religion	0.144	0.139	0	0.690	46

Table 5.8: Pairwise Correlations of Cultural Variables with Country Level Variables

Variable	Hardwork	Ratio Hardw.	Comp.	Ratio Comp.	Freechoice	Ratio Freech.
Hardwork	1.000	-0.122	0.102	0.158	-0.349**	-0.275*
Ratio of Hardwork	-0.122	1.000	0.127	0.263*	0.124	0.204
Competitiveness	0.102	0.127	1.000	0.075	0.054	0.022
Ratio of Comp.	0.158	0.263*	0.075	1.000	-0.074	0.224
Freechoice	-0.349**	0.124	0.054	-0.074	1.000	0.655***
Ratio of Freechoice	-0.275*	0.204	0.022	0.224	0.655***	1.000
Mean Test Score	-0.093	0.232	-0.086	0.453***	-0.173	0.079
Tracking	-0.053	0.042	-0.257*	-0.136	-0.112	-0.197
Pre-Primary	-0.234	0.157	-0.248	0.226	-0.097	0.065
Exp. on Prim. Edu.	0.130	0.162	0.131	0.049	-0.364**	-0.298*
Out-of-school Children	-0.123	-0.117	-0.209	-0.404***	0.222	0.085
Return to Edu.	0.007	-0.248	-0.244	-0.289*	0.186	0.352**
per capita GDP	-0.383***	0.274*	-0.058	0.371***	0.145	0.398***
Gini Index	-0.082	-0.304**	0.009	-0.346**	0.301**	0.282*
Share of Protestants	-0.517***	0.357**	0.117	0.294*	0.328**	0.223
Share of Catholics	-0.410***	0.047	-0.324**	-0.297**	0.222	0.121
Share of Muslims	0.275*	-0.086	0.261*	-0.072	-0.023	0.016
Share of Buddhists	0.099	-0.242	-0.199	0.057	-0.117	0.068
Share of Hindus	-0.022	0.002	0.113	-0.062	0.209	0.261*
Share of no Religion	0.082	0.101	-0.099	0.272*	-0.159	-0.098

Notes: Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

5.C Native Students, Descriptive Statistics

Table 5.9: Student Observations by Country

Country Code	Obs.	PISA score	Country Code	Obs.	score	Country Code	Obs.	score
Central Europe			*Eastern Europe*			*Asia*		
AUT	3,481	506	ALB	4,641	394	CHN	5,075	613
BEL	5,805	515	BGR	4,966	439	HKG	1,789	561
CHE	6,088	531	CZE	4,659	499	JPN	5,968	536
DEU	3,208	514	EST	3,722	521	KOR	4,949	554
FRA	4,613	495	HRV	3,485	471	MAC	841	538
LIE	79	535	HUN	4,426	477	TAP	5,731	560
LUX	1,778	490	LTU	4,127	479	*Southeast Asia*		
NLD	3,423	523	LVA	3,349	491	IDN	5,532	375
Southern Europe			MNE	3,501	410	MYS	4,868	421
ESP	20,643	484	POL	4,424	518	SGP	3,389	573
GRC	4,095	453	ROU	4,977	445	THA	6,467	427
ITA	25,932	485	SRB	3,550	449	VNM	4,906	511
PRT	4,321	487	SVK	4,350	482	*Oceania*		
Britain			SVN	4,772	501	AUS	9,184	504
GBR	9,519,	494	*North America*			NZL	2,216	500
IRL	3,450	501	CAN	15,087	518	*Rest*		
Scandinavia			USA	3,454	481	ARE	4,166	434
DNK	4,706	500	*Latin America*			ISR	3,260	466
FIN	6,340	519	ARG	4,975	388	JOR	5,009	
ISL	2,819	493	BRA	18,391	391	QAT	4,3206	376
NOR	3,563	489	CHL	6,396	423	TUN	4,160	388
SWE	3,235	478	COL	8,747	376	TUR	4,638	448
Russia			CRI	3,914	407	*Total*	373,428	
KAZ	4,503	432	MEX	32,006	413			
RUS	4,061	482	PER	5,842	368			
			URY	4,918	409			

Table 5.10: Summary Statistics of Student Level Variables Included in Equation (5.2)

Variable	Mean	Std. Dev.	Min.	Max.	N
Achievement	453.062	103.832	19.793	962.229	383,394
Parents' Education	10.854	3.863	0	18	373,428
Age	15.8	0.292	15.17	16.33	383,279
Female	0.503	0.5	0	1	383,394
Single Parent	0.125	0.33	0	1	383,394

Notes: Weighted by students sampling probability.

Table 5.11: Pairwise Correlations of Student Level Variables Included in Equation (5.2)

	Achievement	Parent' Educ.	Age	Female
Parent' Educ.	0.391***	1.0000		
Age	-0.0093***	-0.038***	1.0000	
Female	-0.052***	-0.035***	-0.002	1.0000
Single Parent	-0.006***	0.034***	0.003*	0.007***

Notes: Weighted by students sampling probability. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

5.D Second Generation Immigrants, Descriptive Statistics

Table 5.12: Number of Observations by Test and Home Country

Test Country	Origin Country	Obs.	Test Country	Origin Country	Obs.
ARG			HKG		
	BRA	23		CHN	2,716
	CHL	18	HRV		
	URY	29		BIH	711
AUS				SRB	37
	CHN	224	IRL		
	GBR	388		GBR	109
	HKG	44	ISR		
	IND	92		FRA	27
	ITA	12		RUS	62
	KOR	27		USA	49
	NZL	158	KGZ		
	PHL	108		RUS	31
	USA	14		UZB	18
	VNM	157	LIE		
	ZAF	31		CHE	16
AUT				ITA	13
	BIH	262		TUR	16
	DEU	19	LUX		
	POL	38		DEU	41
	ROU	47		FRA	88
	SRB	84		ITA	124
	TUR	433	LVA		
BEL				RUS	160
	DEU	13		UKR	42
	FRA	21	MAC		
	NLD	24		CHN	7,653
	TUR	297		HKG	17
CHE				PHL	42
	ALB	86	MDA		
	DEU	77		UKR	11
	ESP	133	MEX		
	FRA	58		USA	16

Table 5.13: (continued)

Test Country	Origin Country	Obs.	Test Country	Origin Country	Obs.
	ITA	420	MNE		
	TUR	467		BIH	69
CZE				SRB	87
	VNM	63	NLD		
DEU				CHN	24
	BIH	19		MAR	179
	ITA	40		TUR	211
	POL	153	NZL		
	SRB	12		CHN	87
	TUR	445		GBR	85
DNK				KOR	12
	IRQ	87	PRT		
	LBN	16		BRA	12
	PAK	142	QAT		
	TUR	597		EGY	591
FIN				JOR	84
	CHN	12		PSE	441
	EST	39		YEM	639
	IRQ	29	SRB		
	RUS	68		MNE	11
	TUR	16	TUR		
GBR				BGR	19
	CHN	18			
	PAK	31			
GRC			**Total**		**20,300**
	ALB	159			

Table 5.14: Summary Statistics of Variables, Second Generation Immigrants

Variable	Mean	Std. Dev.	Min.	Max.	N
Achievement	481.302	104.646	95.97	890.1	20,300
Parents' Educ. (in years)	11.467	3.998	0	18	20,300
Age	15.764	0.289	15.25	16.33	20,300
Female	0.486	0.5	0	1	20,300
Wealth	-0.118	0.947	-4.96	3.601	20,272
Singleparent	0.071	0.256	0	1	20,300
Hardwork_j	0.599	0.197	0.179	0.927	39
Comp_j	0.531	0.159	0.209	1	39
Freechoice_j	0.497	0.185	0	0.824	39
Comlang	0.388	0.487	0	1	84
Female Employment_j (in %)	45.869	13.994	14.7	72.8	39
Out-of-school Children85_j	12.395	10.649	0.202	40.274	38
Year 2006	0.292	0.455	0	1	20,300
Year 2009	0.321	0.467	0	1	20,300
Year 2012	0.387	0.487	0	1	20,300

Notes: Means of PISA variables are weighted with student sampling weights.

Table 5.15: Pairwise Correlations of Variables Included in Equation (5.3) and (5.4)

	Hardwork_j	Comp_j	Freechoice_j	Dist.Hardw.	Dist.Comp.	Dist.Freech.
Comp_j	0.109	1.000				
Freechoice_j	-0.094	-0.085	1.000			
Dist. Hardwork	0.616***	-0.073	-0.091	1.000		
Dist. Comp	0.128	0.784***	-0.040	0.096	1.000	
Dist. Freechoi.	0.052	-0.089	0.679***	-0.043	0.122	1.000
Comlang	-0.379***	-0.012	0.127	-0.617***	-0.190	-0.028
Femploy_j	0.185*	-0.197*	0.182*	0.125	-0.207	0.180
Outofschool_j	-0.060	0.113	0.038	-0.120	0.219	-0.140

Notes: Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

5.D.1 Male Students

Table 5.16: Second Generation Immigrants, only Males

Variables	(1)	(2)	(3)	(4)
Parents' Educ.	2.008**	6.821*	3.827	2.700
	(0.867)	(4.070)	(4.162)	(4.614)
Parents' Educ. × Hardwork_j		10.123***	16.545***	15.190***
		(3.506)	(4.263)	(4.919)
Parents' Educ. × Comp_j		-11.580***	-13.335***	-10.768**
		(3.763)	(3.918)	(4.812)
Parents' Educ. × Freechoice_j		-11.603**	-14.440**	-14.405**
		(5.432)	(5.740)	(6.001)
Parents' Educ. × Comlang			5.808***	5.434**
			(1.932)	(2.212)
Parents' Educ. × Female Employment_j				0.042
				(0.051)
Parents' Educ. × Out-of-school Children85_j				-0.083
				(0.083)
Student Controls	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Origin × Test Country FE	Yes	Yes	Yes	Yes
Avrg. R^2	0.41	0.41	0.42	0.43
Test Country obs.	23	23	23	23
Origin Country obs.	36	36	36	35
Student obs.	9,989	9,989	9,989	9,807

Notes: Dependent Variable: Math achievement in PISA 2006, 2009 and 2012. Included control variables: age, female, singleparent and wealth. Pooled OLS regression adjusted for student sampling weights. Standard errors are clustered at the test country × origin country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.17: Cultural Distance with Second Generation Immigrants, only Males

Variables	(1)	(2)	(3)
Parents' Educ.	0.781	-2.862	-8.719*
	(1.366)	(2.542)	(4.519)
Parents' Educ. × Dist. Hardwork	0.591	10.360*	11.523*
	(3.230)	(5.975)	(6.138)
Parents' Educ. × Dist. Comp	-6.634**	-5.891*	-3.457
	(3.108)	(3.081)	(4.102)
Parents' Educ. × Dist. Freechoice	-5.479	-3.728	-6.981**
	(3.831)	(3.635)	(3.181)
Parents' Educ. × Comlang		7.272**	7.543**
		(3.177)	(3.158)
Parents' Educ. × Female Employment_j			0.121**
			(0.059)
Parents' Educ. × Out-of-school Children85_j			-0.026
			(0.121)
Student Controls	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Origin × Test Country FE	Yes	Yes	Yes
Avrg. R^2	0.49	0.49	0.51
Test Country obs.	12	12	12
Origin Country obs.	31	31	30
Student obs.	4,337	4,337	4,283

Notes: Dependent Variable: Math achievement in PISA 2006, 2009 and 2012. Included control variables: age, singleparent and wealth. Pooled OLS regression adjusted for student sampling weights. Standard errors are clustered at the test country × origin country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

5.E Robustness Checks

Table 5.18: Impact of Culture using Books as FB Proxy

Variables	(1)	(2)	(3)	(4)
Books	21.931***	13.306	18.336**	-9.333
	(7.091)	(8.291)	(8.463)	(8.800)
Books × Hardwork	-1.908*	0.842	-4.757***	-2.736**
	(1.095)	(1.183)	(1.271)	(1.338)
Books × Ratio of Hardwork	4.888***	-0.521	8.937***	1.535
	(1.417)	(1.720)	(1.692)	(1.639)
Books × Competitiveness	-5.824***	-5.898***	-2.933	-3.798**
	(1.438)	(1.511)	(1.877)	(1.566)
Books × Ratio of Comp.	-5.472	6.086	-13.129*	25.763***
	(6.091)	(6.604)	(7.603)	(7.269)
Books × Freechoice	0.563	5.043***	-4.976**	6.680***
	(1.408)	(1.861)	(2.068)	(1.813)
Books × Ratio of Freechoice	-25.617***	-28.660***	-4.708	-35.427***
	(8.753)	(8.872)	(11.954)	(8.693)
Books × Tracking		0.875***		1.016***
		(0.264)		(0.277)
Books × Pre-primary Enrollment		0.312***		0.247***
		(0.079)		(0.079)
Books × Pre-primary squared		-0.002***		-0.001*
		(0.001)		(0.001)
Books × Exp. on Edu.		0.070*		0.106**
		(0.042)		(0.043)
Books × Out-of-school Children 1985 (%)		-0.088*		-0.051
		(0.052)		(0.052)
Books × Return to Education			0.068	
			(0.121)	
Books × GDP per capita (in 1000 USD)				-0.072***
				(0.014)
Books × Gini Index				0.147***
				(0.050)
Books × Mean Test Score	5.658***	3.000***	3.483***	4.090***
	(0.455)	(0.662)	(0.633)	(0.707)
Student Controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Avrg. R^2	0.43	0.44	0.42	0.43
Country obs.	48	40	35	38
Student obs.	299,260	263,884	254,604	257,266

Notes: Dependent Variable: Math achievement in PISA 2012. Included control variables: age, female, singleparent and people in household. Pooled OLS regression adjusted for student weights. Standard errors are adjusted for clustering at the country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.19: Second Generation Immigrants, FB Measured by Books

Variables	(1)	(2)	(3)	(4)
Books	20.617***	31.058***	30.184***	30.985***
	(1.420)	(7.114)	(6.496)	(7.843)
Books × Hardwork_j		3.397	4.392	6.151
		(6.490)	(6.740)	(7.483)
Books × Comp_j		-16.923***	-17.108***	-16.148**
		(6.304)	(6.490)	(7.052)
Books × Freechoice_j		-11.693	-11.895	-20.914**
		(9.494)	(9.649)	(10.278)
Books × Comlang			1.040	0.038
			(3.116)	(3.405)
Books × Female Employment_j				0.151
				(0.101)
Books × Out-of-school Children85_j				-0.339**
				(0.149)
Student Controls	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Origin × Test Country FE	Yes	Yes	Yes	Yes
Avrg. R^2	0.43	0.44	0.44	0.45
Test Country obs.	29	29	29	29
Origin Country obs.	40	40	40	39
Student obs.	20,850	20,850	20,850	20,486

Notes: Dependent Variable: Math achievement in PISA 2006, 2009 and 2012. Included control variables: age, female, singleparent and wealth. Pooled OLS regression adjusted for student sampling weights. Standard errors are clustered at the test country × origin country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.20: Cultural Distance, FB Measured by Books

Variables	(1)	(2)	(3)
Books	14.639***	16.029***	18.395**
	(3.042)	(4.857)	(7.339)
Books × Dist. Hardwork	22.940***	20.103**	27.641***
	(6.183)	(9.176)	(10.183)
Books × Dist. Comp.	-21.355***	-21.665***	-18.824**
	(7.271)	(7.068)	(7.986)
Books × Dist. Freechoice	-10.489	-10.719	-17.360*
	(9.676)	(9.658)	(10.317)
Books × Comlang		-2.184	0.767
		(4.980)	(4.988)
Books × Female Employment_j			-0.006
			(0.132)
Books × Out-of-school children85_j			-0.532**
			(0.214)
Student Controls	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Origin × Test Country FE	Yes	Yes	Yes
Avrg. R^2	0.50	0.50	0.50
Test Country obs.	15	15	15
Origin Country obs.	34	34	33
Student obs.	8,938	8,938	8,842

Notes: Dependent Variable: Math achievement in PISA 2006, 2009 and 2012. Included control variables: age, female, singleparent and wealth. Pooled OLS regression adjusted for student sampling weights. Standard errors are clustered at the test country × origin country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.21: Regression using Estimated Coefficients of Average Parents' Education as Dependent Variable

Variables	(1)	(2)	(3)	(4)	(5)
Hardwork	-3.177	-2.689	-2.045	-4.067	-2.781
	(2.464)	(2.388)	(2.983)	(2.621)	(2.947)
Comp	-6.109*	-4.858*	-4.981	-2.325	-8.453**
	(3.352)	(2.794)	(4.435)	(4.495)	(3.331)
Freechoice	-7.985**	-6.275**	-6.983*	-12.675***	-6.237*
	(3.240)	(3.030)	(3.821)	(3.767)	(3.386)
Mean Test Score (in 100 points)		2.530***			
		(0.646)			
Pre-primary Enrollment			0.045		
			(0.038)		
Tracking			0.068		
			(1.452)		
Exp. on Prim. Edu. (as % of cGDP)			0.017		
			(0.113)		
Return to Education				0.383*	
				(0.188)	
GDP per capita (in 1000 USD)					0.004
					(0.051)
Gini Index					-0.020
					(0.091)
Out-of-school Children 1985 (%)					-0.222**
					(0.084)
Constant	17.159***	3.520	12.151*	15.599***	19.025***
	(3.521)	(4.797)	(6.098)	(3.813)	(6.413)
R^2	0.26	0.40	0.34	0.40	0.36
Country obs.	45	45	39	35	40

Notes: Dependent Variable: Estimated coefficient on average parents' education per country. This first step regression includes student level variables as in Equation (5.1). First step OLS regression is adjusted for student sampling weights. Standard errors are adjusted for clustering at the school level. Second step regressions are weighted by the inverse of the standard deviation of the first step regression residuals. Robust standard errors are given in parantheses. Albania and Liechtenstein are excluded from the regression due to insignificant first step coefficients. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.22: First Generation Immigrants

Variables	(1)	(2)	(3)	(4)
Parents' Educ.	4.608***	3.249	3.460	0.956
	(0.726)	(3.105)	(3.270)	(3.623)
Parents' Educ. × Hardwork_j		5.475*	5.481	3.642
		(3.259)	(3.399)	(3.366)
Parents' Educ. × Comp_j		-0.212	-0.434	4.062
		(3.092)	(3.007)	(2.874)
Parents' Educ. × Freechoice_j		-3.613	-3.805	-5.255
		(3.378)	(3.487)	(3.785)
Parents' Educ. × Comlang			-0.168	-1.229
			(1.169)	(1.286)
Parents' Educ. × Female Employment_j				0.104**
				(0.048)
Parents' Educ. × Out-of-School Children85_j				-0.200
				(0.129)
Student Controls	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Origin × Test Country FE	Yes	Yes	Yes	Yes
Avrg. R^2	0.37	0.38	0.37	0.39
Test Country obs.	28	28	28	28
Origin Country obs.	38	38	38	37
Student obs.	12,413	12,413	12,404	11,889

Notes: Dependent Variable: Math achievement in PISA 2006, 2009 and 2012. Included control variables: age, female, singleparent and wealth. Pooled OLS regression adjusted for student sampling weights. Standard errors are clustered at the test country × origin country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 5.23: First Generation Immigrants, Cultural Distance

Variables	(1)	(2)	(3)
Parents' Educ.	3.056**	4.201**	2.520
	(1.552)	(1.964)	(3.786)
Parents' Educ. × Dist. Hardwork	2.041	-0.490	-0.862
	(3.081)	(4.112)	(4.126)
Parents' Educ. × Dist. Comp	-7.054*	-7.908**	-2.667
	(3.690)	(3.726)	(2.973)
Parents' Educ. × Dist. Freechoice	-9.099***	-10.530***	-11.592***
	(3.471)	(3.581)	(3.709)
Parents' Educ. × Comlang		-3.022	-2.891
		(2.386)	(2.295)
Parents' Educ. × Female Employment_j			0.074
			(0.053)
Parents' Educ. × Out-of-School Children85_j			-0.197
			(0.123)
Student Controls	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Origin × Test Country FE	Yes	Yes	Yes
Avrg. R^2	0.43	0.43	0.44
Test Country obs.	15	15	15
Origin Country obs.	34	34	33
Student obs.	7,141	7,132	6,957

Notes: Dependent Variable: Math achievement in PISA 2006, 2009 and 2012. Included control variables: age, female, singleparent and wealth. Pooled OLS regression adjusted for student sampling weights. Standard errors are clustered at the test country × origin country level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Bibliography

Abdellaoui, M., H. Bleichrodt, and O. Haridon (2008). A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty 36*(3), 245–266.

Abel, A. B. (1990). Asset prices under habit formation and catching up with the joneses. *The American Economic Review 80*(2), 38–42.

Alesina, A. and P. Giuliano (2011). Preferences for redistribution. In A. B. Jess Benhabib and M. O. Jackson (Eds.), *Handbook of Social Economics*, Volume 1, pp. 93–132. Elsevier.

Alesina, A. and P. Giuliano (2015). Culture and institutions. *Journal of Economic Literature 53*(4), 898–944.

Ammermüller, A. (2005). Educational opportunities and the role of institutions. *ZEW Discussion Papers 05-44*.

Argys, L. M., D. I. Rees, and D. J. Brewer (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management 15*(4), 623–645.

Arnott, R. and J. Rowse (1987). Peer group effects and educational attainment. *Journal of Public Economics 32*(3), 287–305.

Association of Religion Data Archives (ARDA) (2016). World religion dataset. http://www.thearda.com/Archive/Files/Descriptions/WRDNATL.asp.

Aydemir, A., W.-H. Chen, and M. Corak (2013). Intergenerational education mobility among the children of Canadian immigrants. *Canadian Public Policy 39*(Supplement 1), 107–122.

Azmat, G. and N. Iriberri (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics 94*(7), 435–452.

Baumert, J., E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K. Tillmann, and M. Weiß (2001). *PISA 2000: Zusammenfassung zentraler Befunde*. Berlin: Max-Planck-Institut für Bildungsforschung.

Becker, G. S. and N. Tomes (1979). An equilibrium theory of the distribution of income and intergenerational mobility. *The Journal of Political Economy 87*(6), 1153–1189.

Benabou, R. (1996). Equity and efficiency in human capital investment: The local connection. *The Review of Economic Studies 63*(2), 237–264.

Betts, J. R. and J. L. Shkolnik (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review 19*(1), 1–15.

Beugnot, J., B. Fortin, G. Lacroix, and M. C. Villeval (2013). Social networks and peer effects at work. *IZA Discussion Paper 7521*.

Black, S. E. and P. J. Devereux (2011). Recent developments in intergenerational mobility. In D. Card and O. Ashenfelter (Eds.), *Handbook of Labor Economics*, Volume 4, Chapter 16, pp. 1487–1541. Elsevier.

Blanden, J. (2013). Cross-country rankings in intergenerational mobility: A comparison of approaches from economics and sociology. *Journal of Economic Surveys 27*(1), 38–73.

Bock, O., I. Baetge, and A. Nicklisch (2014). hroot: Hamburg registration and organization online tool. *European Economic Review 71*(2014), 117–120.

Brock, W. A. and S. N. Durlauf (2001). Interactions-based models. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5, pp. 3297–3380. Elsevier.

Brunello, G. and D. Checchi (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy 22*(52), 781–861.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica 81*(3), 855–882.

Carroll, C. D., B.-K. Rhee, and C. Rhee (1994). Are there cultural effects on saving? Some cross-sectional evidence. *The Quarterly Journal of Economics 109*(3), 685–699.

Checchi, D., C. V. Fiorio, and M. Leonardi (2013). Intergenerational persistence of educational attainment in Italy. *Economics Letters 118*(1), 229–232.

Chevalier, A., K. Denny, and D. McMahon (2003). A multi-country study of intergenerational educational mobility. *ISSC Discussion Paper Series 2003/06*.

Clark, A. and A. Oswald (1998). Comparison-concave utility and following behaviour in social and economic settings. *Journal of Public Economics 70*(1), 133–155.

Corak, M. (2006). Do poor children become poor adults? Lessons from a cross country comparison of generational earnings mobility. In J. Creedy and G. Kalb (Eds.), *Research on Economic Inequality. Volume 3. Dynamics of Inequality and Poverty*, pp. 143–188. Elsevier.

Corak, M. (2013). Income inequality, equality of opportunity, and intergenerational mobility. *The Journal of Economic Perspectives 27*(3), 79–102.

Cortes, K. E. and J. S. Goodman (2014). Ability-tracking, instructional time, and better pedagogy: The effect of double-dose algebra on student achievement. *The American Economic Review 104*(5), 400–405.

Cox, T. H., S. A. Lobel, and P. L. McLeod (1991). Effects of ethnic group cultural differences on cooperative and competitive behavior on a group task. *Academy of Management Journal 34*(4), 827–847.

Currie, J. and E. Moretti (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *The Quarterly Journal of Economics 118*(4), 1495–1532.

Dohmen, T. and A. Falk (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *The American Economic Review 101*(2), 556–590.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review 101*(5), 1739–74.

Epple, D. and R. Romano (2011). Peer effects in education: A survey of the theory and evidence. In A. B. Jess Benhabib and M. O. Jackson (Eds.), *Handbook of Social Economics*, Volume 1, Chapter 20, pp. 1053–1163. Elsevier.

Falk, A. and A. Ichino (2006). Clean evidence on peer effects. *Journal of Labor Economics 24*(1), 39–57.

Figlio, D. N. and M. E. Page (2002). School choice and the distributional effects of ability tracking: Does separation increase inequality? *Journal of Urban Economics 51*(3), 497–514.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Foster, G. and P. Frijters (2010). Students' beliefs about peer effects. *Economics Letters 108*(3), 260–263.

Fuchs, T. and L. Wößmann (2008). What accounts for international differences in student performance? A re-examination using PISA data. In C. Dustmann, B. Fitzenberger, and S. Machin (Eds.), *The Economics of Education and Training*, pp. 209–240. Springer.

Gali, J. (1994). Keeping up with the Joneses: Consumption externalities, portfolio choice, and asset prices. *Journal of Money, Credit and Banking 26*(1), 1–8.

Galindo-Rueda, F. and A. Vignoles (2007). The heterogeneous effect of selection in UK secondary schools. In L. Woessmann and P. E. Peterson (Eds.), *Schools and the Equal Opportunity Problem*, Chapter 5, pp. 103–128. CESifo Seminar Series.

Gibbons, F. X. and B. P. Buunk (1999). Individual differences in social comparison: development of a scale of social comparison orientation. *Journal of Personality and Social Psychology 76*(1), 129–142.

Gill, D., Z. Kissová, J. Lee, and V. L. Prowse (2015). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *IZA Discussion Papers 9286*.

Gill, D. and V. Prowse (2012). A structural analysis of disappointment aversion in a real effort competition. *The American Economic Review 102*(1), 469–503.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *Qarterly Journal of Economics 118*(3), 1049–1074.

Gneezy, U. and A. Rustichini (2004). Gender and competition at a young age. *The American Economic Review 94*(2), 377–381.

Guiso, L., P. Sapienza, and L. Zingales (2003). People's opium? Religion and economic attitudes. *Journal of Monetary Economics 50*(1), 225–282.

Guiso, L., P. Sapienza, and L. Zingales (2006). Does culture affect economic outcomes? *The Journal of Economic Perspectives 20*(2), 23–48.

Guiso, L., P. Sapienza, and L. Zingales (2009). Cultural biases in economic exchange? *Quarterly Journal of Economics 124*(3), 1095–1131.

Hahn, J. and J. Hausman (2003). Weak instruments: Diagnosis and cures in empirical econometrics. *American Economic Review 93*(2), 118–125.

Hamilton, B. H., J. A. Nickerson, and H. Owan (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy 111*(3), 465–497.

Hannan, R. L., R. Krishnan, and A. H. Newman (2008). The effects of disseminating relative performance feedback in tournament and individual performance compensation plans. *The Accounting Review 83*(4), 893–913.

Hanushek, E. and L. Woessmann (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal 116*(510), C63–C76.

Hanushek, E. A., S. Link, and L. Wößmann (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics 104*(2013), 212–232.

Hanushek, E. A. and L. Woessmann (2014). Institutional structures of the education system and student achievement: A review of cross-country economic research. In R. Strietholt, W. Bos, J.-E. Gustafsson, and M. Rosen (Eds.), *Educational Policy Evaluation through International Comparative Assessments*, pp. 145–175. Waxmann Verlag.

Hayward, R. D. and M. Kemmelmeier (2007). How competition is viewed across cultures. A test of four theories. *Cross-Cultural Research 41*(4), 364–395.

Hertz, T., T. Jayasundera, P. Piraino, S. Selcuk, N. Smith, and A. Verashchagina (2007). The inheritance of educational inequality: International comparisons and fifty-year trends. *The B.E. Journal of Economic Analysis & Policy 7*(2), Article 10.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis 14*(3), 205–227.

Hofstede, G. (1984). *Culture's Consequences: International Differences in Work-Related Values*, Volume 5 of *Cross Cultural Research and Methodology Series*. Newbury Park: SAGE Publications.

Hofstede, G. (1986). Cultural differences in teaching and learning. *International Journal of Intercultural Relations 10*(3), 301–320.

Hofstede, G., G. J. Hofstede, and M. Minkov (2010). *Culture's Consequences: Software of the Mind: Intercultural Cooperation and Its Importance for Survival.* (3 ed.). New York: Mcgraw-Hill Professional.

Horton, N. J. and K. P. Kleinman (2007). Much ado about nothing. *The American Statistician 61*(1), 79–90.

Houston, J. M., P. B. Harris, R. Moore, R. Brummett, and H. Kametani (2005). Competitiveness among Japanese, Chinese, and American undergraduate students. *Psychological Reports 97*(1), 205–212.

Hoxby, C. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics 115*(4), 1239–1285.

Ichino, A., L. Karabarbounis, and E. Moretti (2011). The political economy of intergenerational income mobility. *Economic Inquiry 49*(1), 47–69.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*, Volume 19. Princeton University Press.

Inglehart, R. (2014). World Values Surveys and European Values Surveys, 1981-1984, 1989-1993, 1994-1999, 1999-2004, 2005-2007 and 2010-2014. www.worldvaluessurvey.org.

Isphording, I. E., M. Piopiunik, and N. Rodríguez-Planas (2015). Speaking in numbers: The effect of reading performance on math performance among immigrants. *Economics Letters 139*(2016), 52–56.

Johnson, E., S. Gächter, and A. Herrmann (2006). Exploring the nature of loss aversion. *IZA Discussion Papers 2015*.

Kagan, S. and M. C. Madsen (1971). Cooperation and competition of Mexican, Mexican-American, and Anglo-American children of two ages under four instructional sets. *Developmental Psychology 5*(1), 32.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society 47*(2), 263–291.

Knack, S. and P. Keefer (1997). Does social capital have an economic payoff? A cross-country investigation. *The Quarterly Journal of Economics 112*(4), 1251–1288.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89.

Kuhnen, C. M. and A. Tymula (2012). Feedback, self-esteem, and performance in organizations. *Management Science 58*(1), 94–113.

Lavy, V., O. Silva, and F. Weinhardt (2012). The good, the bad, and the average: Evidence on ability peer effects in schools. *Journal of Labor Economics 30*(2), 367–414.

Levitt, S. D., J. A. List, S. Neckermann, and S. Sadoff (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy 8*(4), 183–219.

Liu, L. and W. S. Neilson (2011). High scores but low skills. *Economics of Education Review 30*(3), 507–516.

Luttmer, E. F. and M. Singhal (2011). Culture, context, and the taste for redistribution. *American Economic Journal: Economic Policy 3*(1), 157–179.

Macdonald, K. (2014). PV: Stata module to perform estimation with plausible values. *Statistical Software Components*.

Manning, A. and J.-S. Pischke (2006). Comprehensive versus selective schooling in England and Wales: What do we know? *IZA Discussion Paper 66*.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies 60*(3), 531–542.

Mas, A. and E. Moretti (2009). Peers at work. *American Economic Review 99*(1), 112–145.

Meghir, C. and M. Palme (2005). Educational reform, ability, and family background. *The American Economic Review 95*(1), 414–424.

Meier, V. and G. Schütz (2007). The economics of tracking and non-tracking. *Ifo Working Paper 50*.

Melitz, J. and F. Toubal (2014). Native language, spoken language, translation and trade. *Journal of International Economics 93*(2), 351–363.

Moffitt, R. A. et al. (2001). Policy interventions, low-level equilibria, and social interactions. *Social Dynamics 4*(45-82), 6–17.

Musset, P. (2012). School choice and equity: Current policies in OECD countries and a literature review. *OECD Education Working Papers 66*.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics 122*(3), 1067–1101.

Nimubona, A.-D. and D. Vencatachellum (2007). Intergenerational education mobility of black and white South Africans. *Journal of Population Economics 20*(1), 149–182.

OECD (2005). *PISA 2003 Data Analysis Manual*. OECD Publishing.

OECD (2010). *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)*. OECD Publishing.

OECD (2012). *OECD Programme for International Student Assessment 2012. Student Questionnaire - Form A*. OECD Publishing.

OECD (2013a). *PISA 2012 Results: What Makes Schools Successful? Resources, Policies and Practices. (Volume IV)*. OECD Publishing.

OECD (2013b). *PISA 2012 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)*. OECD Publishing.

OECD (2014). *PISA 2012 Technical Report*. OECD Publishing.

Oettingen, G. (1995). Cross-cultural perspectives on self-efficacy. In A. Bandura (Ed.), *Self-efficacy in Changing Societies*, Chapter 5, pp. 149–175. Cambridge: Cambridge University Press.

Parente, P. M. and J. Santos Silva (2013). Quantile regression with clustered data. *University of Essex Discussion Paper Series 728*.

Pekkarinen, T., R. Uusitalo, and S. Kerr (2009). School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics 93*(7), 965–973.

Plomin, R., D. W. Fulker, R. Corley, and J. C. DeFries (1997). Nature, nurture, and cognitive development from 1 to 16 years: A parent-offspring adoption study. *Psychological Science 8*(6), 442–447.

Psacharopoulos, G. and H. A. Patrinos (2004). Returns to investment in education: A further update. *Education Economics 12*(2), 111–134.

Roth, G. (2001). *Fühlen, denken, handeln*. Frankfurt am Main: Suhrkamp.

Schafer, J. L. and J. W. Graham (2002). Missing data: Our view of the state of the art. *Psychological Methods 7*(2), 147.

Schütz, G., H. W. Ursprung, and L. Wößmann (2008). Education policy and equality of opportunity. *Kyklos 61*(2), 279–308.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research 60*(3), 471–499.

Solon, G. (2002). Cross-country differences in intergenerational earnings mobility. *The Journal of Economic Perspectives 16*(3), 59–66.

Tabellini, G. (2010). Culture and institutions: Economic development in the regions of Europe. *Journal of the European Economic Association 8*(4), 677–716.

The Economist Intelligence Unit (2012). *The learning curve. Lessons in country performance in education.* Pearson.

Thiemann, K. (2017). Ability tracking or comprehensive schooling? A theory on peer effects in competitive and non-competitive cultures. *Journal of Economic Behavior & Organization 137*, 214–231.

Tversky, A. and D. Kahneman (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–291.

Tversky, A. and D. Kahneman (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics 106*(4), 1039–1061.

UNESCO Institute for Statistics (2016a). Education: Enrolment by level of education. http://data.uis.unesco.org/Index.aspx?queryid=120.

UNESCO Institute for Statistics (2016b). Education: Rate of out-of-school children of primary school age, both sexes (%). http://data.uis.unesco.org/Index.aspx?queryid=120.

Van Veldhuizen, R., H. Oosterbeek, and J. Sonnemans (2014). Peers at work: From the field to the lab. *Tinbergen Institute Discussion Paper 14-051/I*.

VanderHart, P. G. (2006). Why do some schools group by ability? *American Journal of Economics and Sociology 65*(2), 435–462.

Wang, M., M. O. Rieger, and T. Hens (2016). The impact of culture on loss aversion. *Journal of Behavioral Decision Making*, doi: 10.1002/bdm.1941.

Woessmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics 65*(2), 117–170.

Woessmann, L. (2008). How equal are educational opportunities? Family background and student achievement in Europe and the US. *Zeitschrift für Betriebswirtschaft 78*(1), 45–70.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.

World Bank (2014a). Expenditure per student, secondary (% of GDP per capita). http://data.worldbank.org/indicator/SE.XPD.SECO.PC.ZS/countries.

World Bank (2014b). GDP per capita (constant 2005 USD). http://data.worldbank.org/indicator/NY.GDP.PCAP.KD.

World Bank (2015a). GDP per capita (current US$). http://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD.

World Bank (2015b). Government expenditure per student, primary (% of GDP per capita). http://data.worldbank.org/indicator/SE.XPD.PRIM.PC.ZS.

World Bank (2016a). GINI index (World Bank estimate). http://data.worldbank.org/indicator/SI.POV.GINI.

World Bank (2016b). Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate) in 2012. http://data.worldbank.org/indicator/SL.TLF.CACT.FE.ZS.

Appendix A

Summaries

Chapter 2: *Ability Tracking or Comprehensive Schooling? - A Theory on Peer Effects in Competitive and Non-Competitive Cultures*

We develop a model of student decision making that shows that it depends on the culture of competitiveness in a country or region whether it is optimal to choose a school design with ability tracking or comprehensive schooling. Students with different cultural background differ in their concern for relative position in the classroom, which is modeled by reference-dependent preferences. We contrast competitive cultures, where students compare their performance with the best performance in class, and non-competitive cultures where the reference point is the average performance. Taking into account students with heterogeneous abilities, we show that the average performance in competitive cultures is maximized under comprehensive schooling and in non-competitive cultures under ability tracking. Segregation of abilities, however, always leads to a higher dispersion of performances.

Wir entwickeln ein Schüler-Entscheidungsmodell, das zeigt, dass es von der Kultur des Konkurrenzdenkens in einem Land oder einer Region abhängt, ob ein gegliedertes Schulsystem oder ein Gesamtschulsystem optimal ist. Schüler mit unterschiedlichem kulturellen Hintergrund unterscheiden sich in ihrem Interesse an ihrer relativen Position im Klassenzimmer, was wir mit referenzabhängigen Präferenzen modellieren. Wir vergleichen kompetitive Kulturen, in welchen Schüler ihre Leistung mit dem besten Schüler in der Klasse vergleichen und nicht-kompetitive Kulturen, in welchen der Referenz-Punkt die durchschnittliche Leistung ist. Wir betrachten Schüler mit heterogenen Fähigkeiten, um zu zeigen, dass die durchschnittliche Leistung in kompetitiven Kulturen unter einem Gesamtschulsystem maximiert wird und in einer nicht-kompetitiven Kultur mit einem gegliederten Schulsystem. Die Trennung von Schülern mit unterschiedlichen Fähigkeiten führt jedoch immer zu einer größeren Streuung der Schülerleistungen.

Chapter 3: *Does the Impact of Ability Grouping vary with the Culture of Competitiveness? - Evidence from PISA 2012*

In this paper theoretical hypotheses from Thiemann (2017) are tested for their empirical relevance. The theoretical prediction is that comprehensive schooling or ability grouping at the school level yield different results in terms of average student performance in countries that differ in their culture of competitiveness. The predictions are tested using a country-level indicator for competitiveness from the WVS. Educational achievement data is from PISA 2012, covering 34 countries and more than 10,000 schools of which data on the school's policy of ability grouping is available. To overcome possible endogeneity of ability grouping an instrumental variable approach is employed, using the number of schools a school competes with as an instrument. The estimation shows that ability grouping in some or all classes increases average student achievement in competitive cultures and decreases average student achievement in non-competitive cultures.

In diesem Papier überprüfen wir theoretische Hypothesen von Thiemann (2017) auf ihre empirische Relevanz. Die theoretische Vorhersage ist, dass ein Gesamtschulsystem im Gegensatz zu einem gegliedertes Schulsystem zu unterschiedlichen Schülerleistungen führt, je nachdem ob das Land eine kompetitive oder nicht-kompetitive Kultur hat. Die Hypothesen werden unter Verwendung eines Länder-Indikators für Konkurrenzdenken aus dem WVS getestet. Die verwendeten Schülerleistungsdaten stammen aus dem PISA Test 2012 und kommen aus 34 Ländern, mit über 10 000 Schulen, welche sich im Hinblick auf die Gliederung von Schülern in Klassen mit unterschiedlichen Schwierigkeitsgraden unterscheiden. Um ein mögliches Endogenitätsproblem dieser Variable zu beseitigen verwenden wir einen Instrumentenvariablen Ansatz, wobei wir die Anzahl der Schulen mit denen die betrachtete Schule in Konkurrenz steht als Instrument verwenden. Unsere Schätzung zeigt, dass die Gliederung von Schülern die Leistungen in kompetitiven Ländern erhöht, in nicht-kompetitiven Ländern jedoch senkt.

Chapter 4: *An Experiment on Peer Effects under Different Relative Performance Feedback and Grouping Procedures*

We conduct a laboratory experiment to test theoretical predictions from Thiemann (2017) on subjects' performance in an effort task conditional on their peer group's composition and relative performance feedback. Subjects are grouped either randomly or according

to their ability, with the feedback being the maximum or average performance of their group. We are able to support theory-derived hypotheses on optimal performance and peer effects. While no support is found for outcome differences between random and ability grouping, the results show evidence on output being more dispersed when maximum performance is given as feedback. We also find gender differences with respect to peer effects. Male subjects perform significantly better when they compare themselves to the best peer instead of the average, while the opposite is true for females.

Wir führen ein Laborexperiment durch, um theoretische Hypothesen von Thiemann (2017) über die Leistung von Testpersonen unter unterschiedlichen Gruppenkonstellationen und unterschiedlichem relativen Leistungsfeedback zu überprüfen. Die Testpersonen werden entweder zufällig oder entsprechend ihren Fähigkeiten Gruppen zugeteilt, in welchen sie entweder Feedback über die durchschnittliche oder über die beste Leistung in der Gruppe erhalten. Wir können theoretische Hypothesen über die optimale Leistung und Peer Effekte bestätigen. Jedoch finden wir keinen signifikanten Unterschied zwischen zufällig gemischten Gruppen und nach Fähigkeit zugeordneten Gruppen. Wenn die Testpersonen den besten Gruppenwert als Feedback bekommen, sind die Leistungen ungleicher. Wir finden auch Geschlechterunterschiede im Hinblick auf Peer Effekte. Männliche Testpersonen haben signifikant höhere Leistungen, wenn sie sich mit dem Besten vergleichen, während Frauen besser sind, wenn sie sich mit dem Durchschnitt vergleichen.

Chapter 5: *Culture as a Determinant of Intergenerational Education Mobility - Evidence from PISA*

This paper analyzes the determinants of cross-country differences in the impact of family background on student achievement. The focus among potential drivers is on country-specific family culture that can influence student motivation. Measures for culture are derived from questions in the WVS about the valuation of hard work, competitiveness and the belief in free choice in life. In the first part of this paper we focus on native students to compare intergenerational mobility among more than 40 countries. In a second part data from students with immigration background is used in order to overcome endogeneity problems of the cultural variables. We find that disadvantages caused by family background can be overcome more easily in cultures with high beliefs in free choice. Especially male students also benefit if they come from competitive cultures. A high valuation of hard work, however, can decrease mobility.

Dieses Papier analysiert die Determinanten der länderübergreifenden Unterschiede in den Auswirkungen des familiären Hintergrunds auf die Schülerleistung. Der Fokus unter den potentiellen Ursachen liegt auf der landesspezifischen Familienkultur, die die Motivation der Schüler beeinflussen kann. Messzahlen für die Kultur stammen aus Fragen des WVS über die Wertschätzung harter Arbeit, Konkurrenzdenken und den Glauben an Kontrolle über das eigene Leben. Im ersten Teil dieses Papiers konzentrieren wir uns auf Schüler ohne Migrationshintergrund, um die intergenerationale Mobilität in mehr als 40 Ländern zu vergleichen. In einem zweiten Teil werden Daten von Schülern mit Migrationshintergrund verwendet, um die Endogenitätsprobleme der Kulturvariablen zu beseitigen. Unsere Resultate zeigen, dass Nachteile, die durch den Familienhintergrund verursacht werden, leichter in Kulturen mit einem hohen Glauben an Kontrolle über das eigene Leben überwunden werden können. Vor allem männliche Schüler profitieren auch von Konkurrenzdenken. Eine hohe Wertschätzung von harter Arbeit kann jedoch die Mobilität verringern.