The not so widening achievement gap: an interntational comparison of the evolution of the SES achievement gap from 2000-2015

Jorge Cimentada

September 11, 2017

Contents

1 Introduction

2 Literature Review

Educational inequality and its long term impacts is a topic that has been prominent in the social science literature for the past 30 years. The idea of meritocracy and intergenerational transfers has motivated, for a good part of the 20th and 21st century, much of the research on social mobility and social inequalities. When James Coleman released his famous Coleman report (Coleman et al., 1966), it raised the topic to starlight by suggesting that family socio-economic status (SES) and student performance are tightly linked. The topic has been studied extensively since the report was released and several authors have contested whether the relationship is an invariable social law or a product of institutional arrangements. For a detailed review of the long literature on educational inequality, please see Gamoran (2001). But fast-forward until today, we've understood the relationship much better.

Developmental psychology has been studying the subject for a long time and they find that the early stages in a child's life course are extremely important, if not the most, for cognitive development and defining personality traits (Duyme et al., 1999; Waldfogel, 2006). Economists and educators have long neglected these findings but the work of James Heckman has brought our attention to it. For example, he has clearly showed that cognitive and non-cognitive inequalities are present before entering school (Heckman, 2006). This means that inequality starts strong before the schooling of a child begins and schools have the potential to equalize the learning ground. These inequalities can be narrowed significantly with high quality learning environments, specially for children coming from unfavourable conditions. But one of the most important findings from Cunha et al. (2006) is that the cognitive level of a child in time t is a direct function of the experiencies from time $t-1$. Although this sounds logical, it's importance is not understood entirely.

The implications of this cumulative model suggests that it is much less cost-effective to invest in children in time t that in $t-1$. That is why the whole debate about narrowing inequality between high and low SES families should start in early education – it is cheaper than investing at later stages and it brings about the largest returns. Despite all of these findings and strategies to reduce the gaps, we still find that the relationship between children's parental education and future destination is present in virtually all empirical studies of social mobility and inequalities (Breen and Goldthorpe, 1997; Breen and Jonsson, 2007; Waldfogel, 2006; Bradbury et al., 2015; Chetty et al., 2016).

For this reason, it is incomplete to study educational inequalities without considering the prevalence of intergenerational transmission of human and social capital. Social mobility has deepend our knowledge on the relationship between intergenerational transmission of SES and inequality of opportunity. The field of economics of education has also helped to understand that education is by far the investment that yields the best returns. This is not only for the child and his/her family, but to society as a whole, as it boosts economic activity, it helps the labour market improve job conditions and maintain a rapid economic growth (Hanushek and Wößmann, 2007). Since Coleman et al. (1966) we've spent most of our time studying the mechanisms through which this inequality comes about. Naturally, we want to do that in order to reverse it and help every child reach its fullest potential. Despite our good intentions, we've concentrated little on the magnitud of the gaps, specially in terms of cognitive abilities. We have very little evidence, in comparative terms, for which countries have big or small cognitive gaps resulting from SES origins. And even more importantly, we haven't assesed whether policy efforts to reduce inequality have actually had an impact in reducing the achievement gap over time.

However, the closest we have is the vast literature on social mobility. Thanks to these findings, we know that virtually in all countries, be it developed or developing, there is inequality of opportunity. But there is considerable variation in the magnitud of this effect. For example, the Scandinavian

countries, particularly Denmark and Sweden, prove to be very mobile countries (Esping-Andersen and Wagner, 2012; Breen and Jonsson, 2007; Shavit and Blossfeld, 1993).

From Denmark, for example, the literature has learned a great deal about how to improve social mobility. But that was possible by first learning that Denmark is by far one of the most mobile countries, boosting the chances of upward mobility specially for less advantaged children (Björklund and Jäntti, 2009; Jæger and Holm, 2007). The important finding came when we discovered the main reason behind their social escalator: it's educational system. For example, research by Esping-Andersen et al. (2012) and Bauchmüller et al. (2014), show that the Danish early education system has important and longstanding impact on improving opportunities. The education system is completely subsidized for all children, otherwise giving opportunities to children wouldn't have been able to pay. In addition to this, Denmark is recognized as a world-leader in terms of public support for its child-care system as it spends around 2% of GDP, and has among the highest enrollment rates for children under 6 years old (Esping-Andersen et al., 2012). Moreover, it tracks students into different curricula very late compared to other Europen countries (age 16), something which has been linked to less educational inequality (Hanushek et al., 2006). These two traits make the Danish system uniquely effective. The early schooling experience attempts to level children to the same degree and this process is not stopped by different curricular tracks because it starts at around age 16 when cognitive abilities are less malleable. In short, the importance of this finding, is that we should step back and first study whether the effect is there, how big is it, and then proceed to find the causes behind it.

The first attempt to study the evolution of the achievement gap has found that the gap in cognitive abilities between High-SES and Low-SES kids has been widening over the years. The literature on the topic has mainly concentrated on studying the American case (Reardon, 2011) but other international evidence is emerging with a very similar landscape. The United States is usually the case of study because it is the only country where cognitive testing is present as early as 1940. This time-trend has allowed researchers to study achievement gaps in a lengthy period of time. Using this information, Reardon (2011) is the first to investigate the evolution of the cognitive gap and the results are very surprising. Not only has the cognitive gap between the 90th income percentile and the 10th income percentile grown over time, but it has grown faster and to be wider than the highly contested white-black gap (Magnuson and Waldfogel, 2008). The gaps have actually reversed and now a days we find that the income achievement gap is nearly twice as large as the black-white achievement gap (quite the opposite to 20 years back).

Reardon (2011) finds that the increase in the gaps has occured predominantely from the 1970's until the 2000's. In fact, the hard numbers suggest that the gap widened by 40-50%. The author also estimates the rate of change using data as early as 1940 and finds an even higher increase of 75%. Given that the studies before 1970 are less reliable in terms of comparability and sampling design, the author computes all results for both before/after 1970. To provide a definitive answer to the size of the gap, Reardon (2011) concludes that the U.S gap between the 90th and 10th income percentile is at about 1.25 standard deviations in the year 2001. Using longitudinal data, Bradbury et al. (2015) find similar results. Their empirical analysis suggests that for American 14 year olds, the gap is above 1 standard deviation but lower than 1.25. Surprisingly, Duncan and Magnuson (2011) find similarly to the previous studies and confirm a gap of 1.25-1.50 standard deviations. One important drawback of these studies is that they don't present the uncertainty of this estimates. Not necessarily to gauge their statistical significance, but to simply asses how much we can trust their accuracy. It could be that the gap is at 1.25 standard deviations, varying up to 1.75 and down to 0.80. Right now we don't know the upper/lower bound of this estimate, making it difficult to compare when new evidence arises.

Interestingly, this widening of the achievement gap has been paralleled by a growth of income inequality, a very suggestive explanation. Reardon (2011) offers several possible links, with the most reasonable being that family investment patterns have changed so that high income families now invest more resources on their children. This explanation lies in the fact that increasing income became more strongly correlated with other positive family traits related to time allocation and welfare services.

However likely, this is a highly contested topic. A recent report by the Economic Policy Institute finds that the black-white income gap has been growing since 1979, very similar to the increase of income inequality. However, if the achiement gap would be a function of income inequality, we should expect for the white-black achievement gap to be widening rather than narrowing, and those are not the findings given by Reardon (2011) and Magnuson and Waldfogel (2008) [1].

In a follow-up study, Reardon and Portilla (2015) unexpectedly uncovered a new finding: the reversal of the trend. This follow-up study concentrated solely on the kindergarten children in which they took a snapshot of the achievement gap in for 1998, 2006 and 2010. They found that the 90th/10th income gap in readiness closed modestly. Furthermore, using data from fall and spring in the same kindergarden year, they calculated that the gap narrowed at a rate of 0.01 and 0.008 standard deviations per year for mathematics and literacy between 1998 and 2010. They also calculated the same changes for a number of personality traits such as self-control and externalizing behaviour and found similar results. In contrast, Reardon (2011) finds that in a 30-year span the gap was systematicaly increasing at a rate of 0.02, something reasonably close to the previous estimates. Their results not only hold for the income achievement gap, but they also found a decline in the white-hispanic gap (although not for the white-black gap). It should be noted that perhaps the reversal of trend in Reardon (2011) would be obvious if data were available for years after 2000, which is specifically the time that Reardon and Portilla (2015) find the reversal.

The reasons why the authors find a reversal in the trend could be numerous and should be studied very closely. The authors incorporate a number of country-level indicators to explain this change and suggest that the reversal is likely due to the high increase of preschool enrollment. They build on their previous argument by suggesting that in this same period (1998 - 2010) the income achievement gap in early schooling enrollment decreased substantially. Their conclusions, although suggestive, are very speculative and have no empirical support, which is why this is still an open question.

Motivated by these recent results, other authors have taken this analysis to an international context in order to discover between-country trends. The work of Bradbury et al. (2015) employs a unique comparative analysis of the achievement gap between Australia, United Kingdom, United States and Canada. Their research design is very distinctive in that they use longitudinal data from children as early as age 2 and study the evolution of the achievement gap up until age 14 [2]. The core finding behind their study is that the American achievement gap is much wider than in any other comparison country, specially Australia and Canada. They find that once the achievement gap is present in early school entry, it doesn't seem to narrow or widen very much over the life course. In fact, they calculate that the quality of early education can only explain about 30-40% of the high school SES gap. This suggests that once the achievement gap is present before entering school, it carries a social-scar effect [3]. One exception is the United Kingdom, which they found to be a country that helps close the gap in early primary years. This can likely be due to the comprehensive schooling and also the public support by the welfare state in dimensions like health and income support.

One limitation of their study is that they concentrate on countries which have a very particular educational structure, namely the fact that there is little formal stratification in terms of curricula. These four countries have no major jump in school selection, quite the opposite to the average European country. A thorough review by Van de Werfhorst and Mijs (2010) sheds some light on the subject. First and foremost, they gather substantive evidence showing that countries which have a highly tracked curriculum tend to have high levels of inequality, measured in terms of achievement gaps. Hanushek et al. (2006), use the Progress in International Reading Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMSS) and the Programme for Inter-

[1] The black-white achievement gap narrowed significantly for cohorts betwen 1950 and 1970 where income inequality was increasing.

[2] To the best of my knowledge this is not only the first study that uses panel data to study achievement gaps, but to also do it between countries

[3] However, schooling could be preventing the gap to widen even more, and very rigorous RCT's show that high quality schooling can indeed help ease the gap, in some instances even close it (Campbell et al., 2002).

national Student Assessment (PISA) surveys to gauge whether highly tracked countries do indeed increase inequality after students pass the age at first selection of tracking. The results suggest that early tracking increases educational inequality. While less clear, there is also a tendency for early tracking to reduce mean performance. Micklewright and Schnepf (2007) using PISA but a different empirical strategy, find that countries which have a high level of tracking, are distinctively unequal in the difference between the top 95th and bottom 5th performers. In fact, the difference in test scores between these two groups is about 10 times higher then the average annual gain of a year of schooling.

Another limitation of the work of Bradbury et al. (2015) is that their analysis is based on four surveys that have significant differences in terms of questions, sampling and populations and cannot be easily compared. They do an amazing work at making the four surveys comparable and their findings are indeed very reliable. But we should be careful at intepreting these findings causally and should be taken as suggestive at best. For this reason we should also pay particular attention to studies such as Chmielewski and Reardon (2016) and Chmielewski (2016) which have attempted not only to compare gaps between countries, but to evaluate whether there is a general increase in the gap using much more comparable surveys. However it should be noted that these studies tackle a completely different question from the above, namely to study cross-sectional differences between many countries, instead of over-time analysis of student gaps. Nonetheless, they do provide support for the overall finding that the achievement gap is certainly not narrowing over time.

Chmielewski and Reardon (2016), again using PIRLS, TIMSS and PISA, assess whether there are patterns of cross-national variation in the achievement gap. In other words, does the achievement gap differ between countries? Their work suggests that there is considerable variation in the achievement gap between top and bottom earning families across many developed countries. In comparison to the literature on achievement gaps, they find that the U.S has a gap of 1.20 SD in 2001 which increase to around 1.30 in the year 2006 while Germany has a decreasing gap from around 1.25 to 1 standard deviation in the same year-span. However, these numbers vary a lot and carry a great deal of overlapping uncertainty.

They go even further and link this achievement gap to several country-level indicators related to income inequality, school differentiation and central exams, among other things. The correlations are very suggestive as explanatory mechanisms but they are very cautious in drawing causality, specially because they are all reasonably correlated among each other.

One interesting question that is still missing from the literature is whether these country gaps have evolved over time. With their data, they only have 3 countries which are present in all waves but they also have very few waves because their question of interest (income categories) was only asked in three time points. That is why their study is more about between country gaps rather than the magnitude and evolution of the gaps.

Chmielewski (2016), building on the work of Chmielewski and Reardon (2016) and Reardon and Portilla (2015) pooled together all the previously mentioned data, together with over 10 more studies ranging from the year 1964 until 2015 in order to discover differences between and across countries. With over 50 years of data, and over 100 countries, she finds that there seems to be a general pattern of increasing achievement gap. However, once she disentangles the relationship by country, she finds a sizable amount of heterogeneity, with some countries seeing the achievement gap narrowing, others no change at all, while others record a steady increase. This is revealing because it doesn't really pay off to look at a general average once each country has their own distinctive gap and evolution. This highlights the notion that the increasing achievement gap is clearly not universal and should be studied in context.

One limitation of their study (as well as Reardon (2011)) is that they adjust for the age of each child in all studies. Although for their modeling purposes this is the correct thing to do [4], these

[4]The differences in achievement could simply be due to changes in cognitive abilities across the lifetime. However, as

modeling strategies are masking age-specific achievement gaps by controlling for age. We clearly see in Reardon and Portilla (2015) that there are age-specific gaps, and they do change at a fast pace in very little time.

In fact, the evolution of high/low SES gaps for preschool children might be much less marked than the same gap for high school children. The explanation, although very debated, has been gaining much support in recent years. If we remember correctly, for countries with high levels of curricular differentiation the transition from early schooling into the tracking system has been found to increase inequality of learning (Hanushek et al., 2006). Moreover, the vast sociological literature on educational transitions systematically finds that tracking tends to foster between-track inequality rather than erode their differences by tackling their specific needs (Van de Werfhorst and Mijs, 2010). Based on this, we cannot simply assume that the achievement gap has been neither constant across cohorts (because there have been tracking reforms in many countries, introducing as well as eliminating tracking structures) nor the same between ages, because tracking/no tracking might exarcebate the achievement gap.

3 Research framework

3.1 Research questions

This study seeks to study the evolution of the high/low SES achievement gap in the past 15 years for all PISA participant countries and search for a likely explanation. This is different from previous work because it concentrates solely on 15 year old children, something not done for the achievement gap, and it attempts to capture the evolution of the achievement gap for countries with different educational systems. In addition, it extends the previous findings by studying the relationship between the gap and the average performance of each country in order to detect if there's an over-time tradeoff between good performance and equality over time separately for each country. The advantages of this study are threefold.

First, we concentrate on the evolution of the gap for only 15 year olds, which will serve as a comparison to the single year-country snapshot of Chmielewski and Reardon (2016) and the evolution of the kindergarten gap in Reardon and Portilla (2015). As we've seen before, there are reasons to think that specific age-groups have seen changes in the achievement gap. Moreover, in almost all countries with a tracked curriculum children are either at or in the process of tracking by the age 15, meaning that we will be able to link whether tracked countries are the most variable in their evolution of achievement gaps.

Second, we will be able, for the first time, to study the evolution of the achievement gap for several countries other than the United States. This will help asses the mangitude of the inequality as well as the rate of growth/decline. This study not only documents the size and changes of the SES gap, but pays particular attention to the source of the changes. That is, we study whether bottom performers are falling behind, if top performers are gaining advantages, or if both phenomenas are simultaneously at play.

Third, we attempt to model why the achievement gap is changing over time. In contrast to other studies, this is a novelty given that most research has been done cross-sectionally between countries.

More formally, we're interested in studying a) Which countries are experiencing changes in the achievement gap; b) establishing whether the gaps are widening/narrowing because particular groups are getting ahead/behind; c) The size of the achievement in a comparative perspective and its relationship to overall achievement levels; d) What explains this cross-country evolution;

we've noted before, Bradbury et al. (2015) find that the achievement gap is very stable across the life time

We develop each question separately for more detail.

a) The seminal work of Reardon (2011) suggests that achievement gaps change, and they do so much quicker that we though. After recording a SES gap increase of about 40% in only 30 years, Reardon and Portilla (2015) stress that they also found a significant decrease in only 15 years of data, showing how important it is to study the changes in the achievement gap over time. We will compare the percentage change at which the gap widened/narrowed from the first to the last year available. This will give us a general idea of the overall change over time, and will allow us to compare our estimates to the actual literature [5]

b) The widening/narrowing of the achievement gap has a source, which has been often studied to be related to everything from educational spending, income inequality, time allocation to students and preschool enrollment. The literature has concentrated very narrowly on whether the gap is increasing because the top performers are getting ahead, because both are distancing or because the bottom is falling behind. We shall pay particular attention to identifying the rate at which the top/bottom groups are evolving over time.

c) We want to investigate whether better performing countries have lower levels of inequality than other countries. Van de Werfhorst and Mijs (2010) emphasize that there is empirical evidence that suggests this. This pattern is not so obvious, however. For example, countries with high leves of tracking could maximize student performance, specially the high SES students, raising their overall performance and thus raising the national performance score. But if the bottom performers are not gaining at the same rate, then the achievement gap will inevitable grow resulting in a high performing countries with a widening achievement gap.

d) Tightly linked to the previous point, we want to test whether several dimensions of the tracking setup explain why the gaps are changing. This will support the previous points in suggesting that the dynamics of the over-time achievement gap are endogenously intertwined with the effect that tracking has on the specific changes of the high and low SES groups.

4 Hypothesis

Following the previous questions and the literature on educational inequality and tracking we pose three hypothesis for answering questions (b), (c) and (d) given that question (a) is purely an exploratory exercise.

- Hypothesis 1: In countries where there is a high degree of tracking we should expect the top and bottom to be evolving at the rate given that tracking is thought to maximize the learning experience of both groups. For countries with low levels of tracking we should expect the top and bottom groups to be closer to each other and for the gap to be closing since the both groups will be in the same track (this doesn't mean that the top is loweing their score but that the bottom is most likely catching up).
- Hypothesis 2: Building on Van de Werfhorst and Mijs (2010), then there should be a negative relationship between the average country performance and their level of inequality. Countries with high performance should have less inequality and countries with low performance should have greater inequality
- Hypothesis 3: if any (or both) of the previous hypothesis are confirmed, then tracking should play an important role on the evolution of the achievement gap. We hypothesize that the degree

[5] Although no study has performed this age-specific achievement gap for comparable tests over such a long time. Our results will serve as comparison for other studies that use age-specific groups, such as 4th graders using PIRLS.

of tracking is tightly related to changes in the gap, and the more tracking, the more inequality. Moreover, the more vocational tracking, the less inequality considering that it gives short term returns in terms of labour market opportunities.

5 Methods

5.1 Data

To investigate the above mentioned questions I will be using the Programme for International Student Assessment (PISA). PISA is a survey carried out every three years that aims to evaluate education systems by testing the skills and knowledge of 15-year-old students. Currrently, PISA has six waves starting in 2000 up until 2015, where recently, over half a million students were tested in mathematics, literacy and science in over 70 developed/developing countries.

PISA collects data through a two-stage stratified sampling design. With the help of official governments, PISA randomly chooses 150 schools in each country, where they then randomly pick thirty 15 year olds to undertake the two hour tests. The sample size for PISA 2000 is 127,388, 276,165 for PISA 2003, 398,750 for PISA 2006, 515,958 for PISA 2009, 480,810 for PISA 2012, and 519,334 for PISA 2015. Together with the subject tests, PISA collects personal information from students, their families and their school environment (including teacher surveys), that serves as relevent background information that can be matched to the students performance. With the recent inclusion of PISA 2015, these six waves make up a time-series analysis of 15 years, enough to visualize changes in the structure of an educational system. None of the literature cited so far has used the last PISA wave, which was recently released in December 2016.

To identify a family's socio-economic status PISA collects several variables that measure different dimensions. Classically, they ask student's their parent's educational level. Scholars have considered this to be a reliable recall given that we expect fifteen year olds to know their parent's level of education (Reardon, 2011). This question has been asked in every wave and holds a somewhat similar coding across time, although the first two waves have small differences. In spite of this, another limitation is the fact that parent's education is measured using the ISCED classification, something that has changed over time. For example, until PISA 2009, the preferred framework was ISCED 1997, whereas the next wave switched to the newly developed ISCED 2011 classification. Both these classification schemes have equivalent look-up tables, but this requires a detailed inspection of the codings.

Another social background variable is the International Socio-Economic Index of Occupational Status (ISEI). This variable attempts to capture the social status of the family, without asking for income information. This index variable was developed by Ganzeboom and Treiman (1996) and later refined by (Ganzeboom, 2010) and it attempts to measure occupational status using a continuous measure. This indicator is a very reliable alternative to the classical Erikson-Goldthorpe-Portocarero classification (Erikson et al., 1979). It has been scaled for comparability between waves and some authors have used it for inequality studies, finding expected results to be consistent with the social mobility literature (Chmielewski, 2016). PISA also includes a plethora of indicators on family wealth, home educational resources, the number of books in the home, among many other material resources in the household.

Yet one of the most relevant variables for our study is a composite SES index created by the PISA team. The index of economic, social and cultural status (ESCS) was created on the basis of the following variables: the International Socio-Economic Index of Occupational Status (ISEI), the highest level of education of the students parents, the PISA index of family wealth (which measures the material wealth of the family), the PISA index of home educational resources; and the PISA index of possessions related to "classical" culture in the family home (mainly about books in the household)

(OECD, 2002). This variable, aside from capturing all relevant dimensions of SES, such as education, occupation, and material resources, takes care of transforming all mentioned variables into comparable metrics across waves.

The ESCS was derived from a principal component analysis of standardised variables, taking the factor scores for the first principal component as measures of the PISA index of economic, social and cultural status. All countries and economies (both OECD and partner countries/economies) contributed equally to the principal component analysis, while in previous cycles, the principal component analysis was based on OECD countries only. However, for the purpose of reporting the ESCS scale has been transformed with zero being the score of an average OECD student and one being the standard deviation across equally weighted OECD countries (OECD, 2016).

To the best of our knowledge, this is the first paper that uses the newly-released ESCS index (OECD, 2016) which was rescaled so that all ESCS indexes are suitable for over-time analysis [6] In other words, the ESCS index does not need any transformation or coding updates because it is ready for comparison over time.

Aside from SES, the other most relevant variables are test scores for mathematics and literacy. PISA does not provide a single test result for each respondent. Instead, it provides a *series* of 'plausible values' that the child could actually score. As explained in the PISA manual (OECD, 2012), these are imputed values that resemble individual test scores and have approximately the same distribution as the latent trait being measured (the true distribution of the possible scores a student can achieve).

A more intuitivity is explanation is this: suppose we have μ_i, the average student test score in mathematics for student i. Instead of estimating μ_i alone, plausible values estimate a distribution of possible μ's for student i, together with the likelihood of each μ_i based on the respondents answers on the test. This is defined as the posterior distributions of μ's for student i. The reason why we use this procedure is because estimating a single estimate μ_i is plagued with measurement error, among other types of bias (see Wu, 2005). The number of plausible values for PISA waves are usually five (although ten for PISA 2015) random draws from this distribution. In practice each student has 5 scores for each test, that resemble their distribution. Those values are continuous, ranging from 0 to 500, with a mean of 250.

5.2 Data analysis

The aim of this paper is to identify, disaggregate and explain country trends in the achievement gap for several countries. To represent the SES gap, most of the literature on achievement gaps has concentrated on indicators such as parental education, parental occupational status, income achievement gaps and actual SES achievement gaps (Fryer and Levitt, 2004; Hanushek et al., 2006; Saw, 2016; Bradbury et al., 2015; Byun and Kim, 2010). The actual calculation of the achievement gap varies substantially and different strategies have been implemented. For example, Micklewright and Schnepf (2007) calculates the difference in achievement by crudely subtracting the gap between the 95th and 5th percentile of the mathematics distribution. Although in principle you should be able to capture some type of SES effect like this, theoretically, it should be much more accurate to difference out the mean score of, for example, parental education or some other SES proxy. Saw (2016), for instance, used parental education as a proxy of SES, whereas Byun and Kim (2010) use a similar SES index as ours, but created by them.

Reardon and Portilla (2015), Chmielewski and Reardon (2016) and Chmielewski (2016) used a different method developed by Reardon (2011) which we partially adopt in this paper. SES achievement gaps are measured as the difference in standardized achievement between the 90th and 10th percentiles of the chosen SES variable. The rule of thumb to choose the 90th, 50th and 10th percentile

[6] These rescaled indices can be found in the PISA website under *Rescaled Indices for Trend Analyses.*

is arbitrary, as others have used, for example, the 95th, 50th and 5th (Micklewright and Schnepf, 2007). We use the conventional 90/10 cutoffs in the literature following the standard set by Reardon (2011).

For each country in each wave, SES disparities in achievement are measured as the gap in standardized achievement between the 90th and 10th percentiles of each countrys distribution of each SES variable, following Reardons (2011) method for income achievement gaps. The original strategy of Reardon (2011) is as follows: first, achievement is standardized (see below for a mathematical explanation of the standardization). We then use it to calculate the mean achievement (and standard error) for each category of the SES variable of interest (parent's education, income categories, etc..). "Category means are plotted at their percentile ranks and cubic models are fit through the points using weighted least squares. Finally, achievement at each countrys 90th and 10th SES percentiles is interpolated from the model" (Chmielewski, 2016). The result is a SES gap from an ordinal variable of interest.

As mentioned before, PISA does not provide a single achievement indicator. Instead, we take the median of all plausible values for each student [7], resulting in one single score.

To standardize the test score we fit a linear model

$$Y_i = \alpha + \beta_1 * AGE_i + \epsilon_i, \quad \epsilon_i \ N(0, \sigma^2) \tag{1}$$

for each wave, where Y_i is the median student test score and AGE_i is their age measured in months (following the same strategy as Reardon (2011) [8]) weighted by the student sample weights.

We then adjust $\hat{\gamma}_i$ by

$$\hat{\gamma}_i = \frac{\hat{\epsilon}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}} \tag{2}$$

Where $\hat{\epsilon}_i$ is the residual for student i and the denominator is the root mean square error of the model.

This new standardized variable has a mean of zero. Standardizing the median test score solves the problem of comparability of gaps measured with different tests, and across waves because the test scores have now the same metric across time. However, if the variance of academic achievement changes over time, then standardizing the overall score at each country-wave pair actually makes the transformation biased. That is, by standardizing we're forcing the standard deviation of test scores to be zero across all waves. But if the true deviations of the median academic achievement changes over time, then the estimated trend in the SES gaps will be underestimated, or vice versa.

We plot the standard deviation of the mathematics test score for all waves in figure 1. The plot suggests that it's something we shouldn't be deeply concerned with. The standard deviation of each wave seems to be following a very similar pattern with a not so drastic exception of the year 2000.

Another concern is that if the gaps at different waves are measured with tests that have different amounts of measurement error, then the amount of bias will not be the same in each measure of the gap. This can be very misleading and suggest erroneous intepretations regarding trends in the of the

[7] Since each plausible value is a random draw from a theoretical latent normal distribution of possible student achievement scores, the median should be precise in getting a central measure of the latent distribution.

[8] This does not mess up our analysis by masking age-specific gaps because all students in the sample are 15 year olds. Control for aging is simply to control for monthly differences in ages.

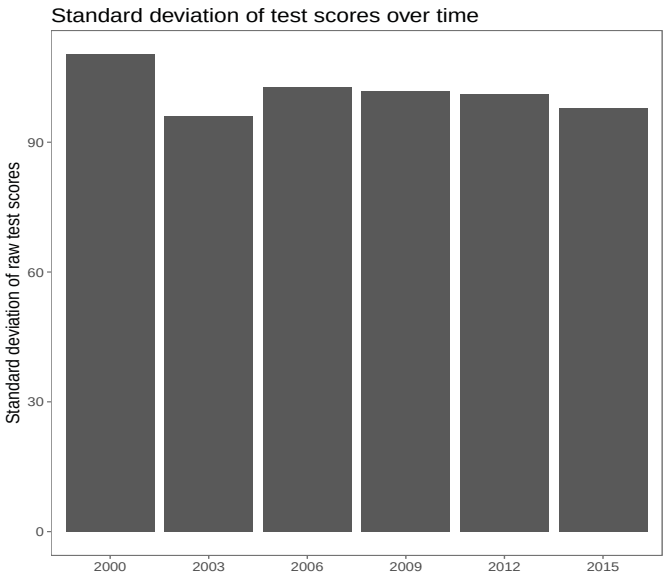

Figure 1: Standard deviation of test scores across all waves

gaps over time (Reardon, 2011). PISA has tried to make sure the tests are comparable across waves (OECD, 2012), but we still have to adjust for this imprecision.

In order to correct the gap estimates for measurement error, we adjust student test scores($\hat{\gamma}_i$) by

$$\hat{\gamma}_i = \hat{\gamma}_i * \frac{1}{\sqrt{r}} \tag{3}$$

where r is the reliability score of the wave [9]. Each PISA survey provides a reliability indicator which we use accordingly. This yields estimates of the true gaps, and eliminates any bias in the trend that may arise from differential reliability of the tests.

After constructing the adjusted test score measurement, we estimate the SES dummy by calculating the weighted 90th and 10th quantiles for the SES composite index, and then generating a dummy of 1 for those above (including) the 90th percentile and 0 for those below (including) the 10th percentile.

We then fit

$$\gamma_i = \alpha_{j[i]} + \beta_{j[i]} * SES_i + \epsilon_i, \text{ for } i = 1, 2, ..., n \text{ for each country } j \tag{4}$$

where SES_i is whether the student is in the top/bottom SES dummy and we allow both α_i and β_i to vary by country j. We implement this separately for each wave and weight by the wave-specific student sample weights.

Finally, we calculate the SES achievement gap for each country by extracting the fitted α and β for each country j and calculating the difference of the predicted score of high and low SES.

[9] Other proceedures multiply each country by their own reliability measure for each year-subject pair (Chmielewski, 2016). The realibility estimates are calculated using Item Response Theory (IRT) analogues of traditional estimates of person separation reliability such as internal consistency. Unfortunately, PISA 2000 did not provide any reliability measure separately for each country and PISA 2015 has to yet release their own. At the moment of writing this paper, they were unavailable. For this reason we implement the analysis following the original work of Reardon (2011)

$$\begin{aligned} \text{High SES}_j &= \alpha_j + \beta_j \\ \text{Low SES}_j &= \alpha_j \\ \text{SES gap}_j &= \text{High SES}_j - \text{Low SES}_j \end{aligned} \tag{5}$$

We also calculate the standard error of this difference and use it to create uncertainty intervals in the estimation of time trends. We fit a multilevel rather than a linear model because by pooling the information together, we weight countries appropriately to their sample size. Given that including the SES *dummy* reduces the sample size considerably, we want to be able to estimate each country-difference as accurately as possible. For further exploration, we also compute the 90th/50th and 50th/10th SES gaps following the same method outlined above.

6 Analysis and results

6.1 Evolution of the achievement gap

The first table shows a description of the sample size and mean score of both top and bottom SES groups for only the first and last time point. One main concern from the planned analysis is that getting the top 90th percentile and bottom 10th percentile would result in a small sample size. The table suggests that we have a reasonable number of respondents to actually estimate gaps accurately. Moreover, we can see that in all instances the bottom SES group has a lower score than all top SES groups.

We see that in some countries like Finland and Sweden the average low SES scores is actually above the average score of 0, suggesting they're very equal countries. However, most countries, like Spain, Poland, Hungary and Germany, their average low SES scores are much lower than the country average, which is 0. We also see some countries with major changes from year 2000 to 2015, with, for example, Australia decreasing the low SES gap from 0.22 standard deviations to -0.15, well below the country average. In the next section we take a look at this in a more detailed fashion, considering all years in between.

We start by look at the achievement for some countries in figure 2. We plot only mathematics for each country and also a quadratic trend spline for *both mathematics and literacy* pooled. Some countries have increased their achievement very strongly. For example, France, Austria and surprisingly Sweden have very steep slopes. France increased the gap by roughly 0.9 standard deviations, Austria by 0.6 and Sweden by 0.6. This pattern happens similarly for literacy. For example, France has an increase of 0.6. For such a short period of time, the magnitud of these increases are reasonably big.

Given that no one has estimated the evolution of the gap we can't cross-check how other empirical estimations put France at. However, we can take Micklewright and Schnepf (2007) as the closest reference which also finds that France was a low dispersion country in 2000. However, there's no evidence on what happened over time. Luckily, the work of Bernardi and Ballarino (2016) did study the relationship of social origin effects (broadly speaking, not in terms of achievement gaps) in France and found that they increased in the last decades.

Other countries have reasonable increases such as Finland and Hungary, with increases of nearly 0.6 and 0.4, respectively. Aside from these countries, there are other cases that have no change at all, specifically, Canada, Netherlands and Spain. Canada excels here not only because the gap has been stable over time, but because it has the smallest gap of all countries presented here. It's nearly 0.5 standard deviations in 2000 and it increased only by 0.2 in 2015. We can't really say Italy has had a significant increase since it has a sort of wiggly pattern that looks like it cancels itself out.

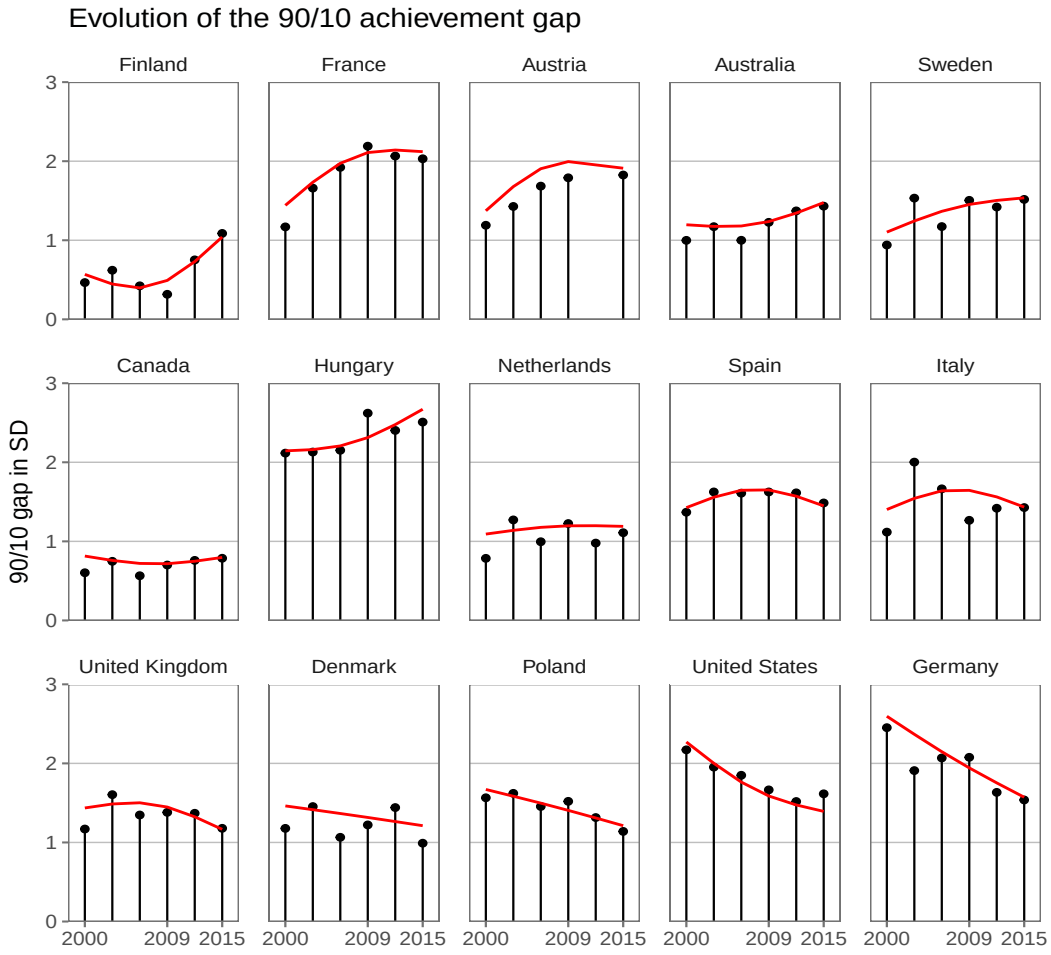

Figure 2: Evolution of the achievement gap for selected countries

		Low SES			High SES		
Year	Countries	N	Avg score	S.E	N	Avg score	S.E
2000	Australia	294	0.22	0.07	270	1.22	0.09
2000	Austria	294	-0.04	0.11	235	1.15	0.15
2000	Canada	1970	0.30	0.05	1357	0.90	0.07
2000	Denmark	230	-0.01	0.13	228	1.17	0.17
2000	Finland	261	0.42	0.12	267	0.89	0.15
2000	France	253	0.04	0.04	258	1.21	0.05
2000	Germany	250	-0.64	0.04	303	1.81	0.05
2000	Hungary	291	-0.46	0.09	254	1.66	0.12
2000	Italy	278	-0.31	0.05	278	0.81	0.06
2000	Netherlands	129	0.46	0.08	158	1.25	0.11
2000	Poland	209	-0.31	0.04	166	1.25	0.06
2000	Spain	318	-0.27	0.05	340	1.09	0.07
2000	Sweden	241	0.12	0.10	246	1.06	0.13
2000	United Kingdom	531	0.16	0.04	415	1.33	0.06
2000	United States	222	-0.43	0.02	164	1.74	0.03
2015	Australia	1694	-0.15	0.04	1235	1.28	0.05
2015	Austria	676	-0.30	0.07	705	1.53	0.10
2015	Canada	2215	0.20	0.03	1863	0.99	0.05
2015	Denmark	1013	0.09	0.08	595	1.08	0.10
2015	Finland	575	0.07	0.08	584	1.16	0.11
2015	France	570	-0.37	0.02	615	1.66	0.03
2015	Germany	545	-0.09	0.02	582	1.45	0.03
2015	Hungary	466	-0.72	0.07	589	1.79	0.09
2015	Italy	947	-0.20	0.03	1043	1.23	0.04
2015	Netherlands	521	0.13	0.04	525	1.24	0.06
2015	Poland	446	0.10	0.03	448	1.24	0.05
2015	Spain	608	-0.25	0.03	703	1.24	0.04
2015	Sweden	527	-0.20	0.06	542	1.32	0.09
2015	United Kingdom	1387	-0.03	0.03	1388	1.15	0.04
2015	United States	585	-0.39	0.01	544	1.22	0.01

Table 1: SES sample size and ISCED composition

On the other hand, we do have some countries which show a decrease in the SES achievement gap. Poland decreased by about -0.4 and Denmark by -0.2. However, the most notable cases are the United States and Germany. These two countries show a very high level of dispersion in the year 2000 with SES gaps of over 2 standard deviations. But in the 15-year time trend both countries reduced the gaps by -0.6 and -0.9 respectively. Their distinctively big gaps in 2000 also show up in the work of Micklewright and Schnepf (2007). This corroborates the findings of Reardon and Portilla (2015), which found a decreasing gap for kindergarteners. This decline is evident also for 15 year olds, suggesting it might be more of an institutional change rather than a specific grade-level policy. Do note that the trendline is for both mathematics and literacy. If it were only for mathematics, then the trendline would probably be even more extreme. Despite this, pooling both subjects gives even more strength to the analysis given that the results stand for both subjects together [10]. We also plot the 80/20 and 70/30 SES gaps and find that very similar patterns as in the 90/10 SES gap (see figure 10 in the appendix).

Analyzing this graph we might naively conclude that these trends are not very steep and they should not be very important to consider. However, remember that the Y axis is measured in standard

[10]Same graph is availabe for reading upon request, where we find pretty much the same results if not even more steep/downward slopes.

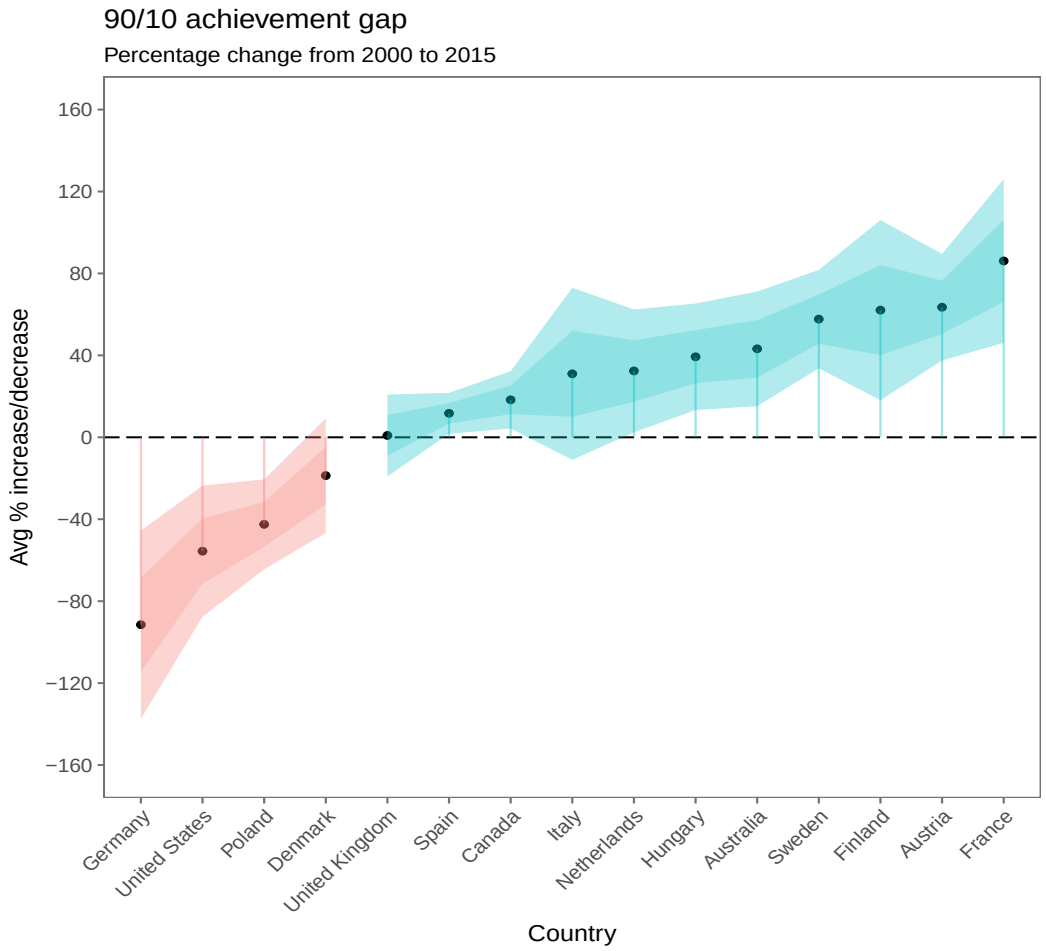

Figure 3: Evolution of the achievement gap for separate SES groups

deviations. Small changes are actually very huge in practical terms. Take the case of Sweden. The slope doesn't look that steep but in reality it increased the gap from 1 standard deviation in 2000 to around 1.5 standard deviation in 2015. Now that we know that, the trends of Poland, United States, France and Germany are particularly shocking.

We find that the initial gap for the U.S in 2000 is 2.17 varying between 2.13 and 2.21. This gap is much higher than found in previous studies, but again, we don't know how much previous estimates vary and nobody has really studied the 15-year old achievement gap precisely. All country-level studies pool many studies together and control for age, removing the age-specific gap effect.

Figure 3 below takes a more direct approach and looks at the percentage change from the first and last timepoint available. Each data point has been computed together with it's 50% and 95% uncertainty interval [11].

Generally speaking, we see that most countries increased their achievement gap over time. France had an average increase of about 80% since 2000 varying down to 40%, whereas Germany had a similar figure but decreasing. Many of the countries we saw before that didn't have a very steep slope, such as Hungary or Australia, in fact had increases of about 40% of their gap. In contrast, we see that the U.S and Poland had also quite significant decreases of about 40%. The benefit of presenting these estimates this way is that we can actually asses the uncertainty of each calculation, and we do find that some of these estimates have wide variability. Despite this, most countries show a clear sign of either decreasing or increasing. We plot the same graph for the 80/20 and 70/30 SES gaps in the appendix and the results hold for these achievement gaps as well.

[11] Each of these uncertainty intervals were computed using a 500-replicate bootstrap

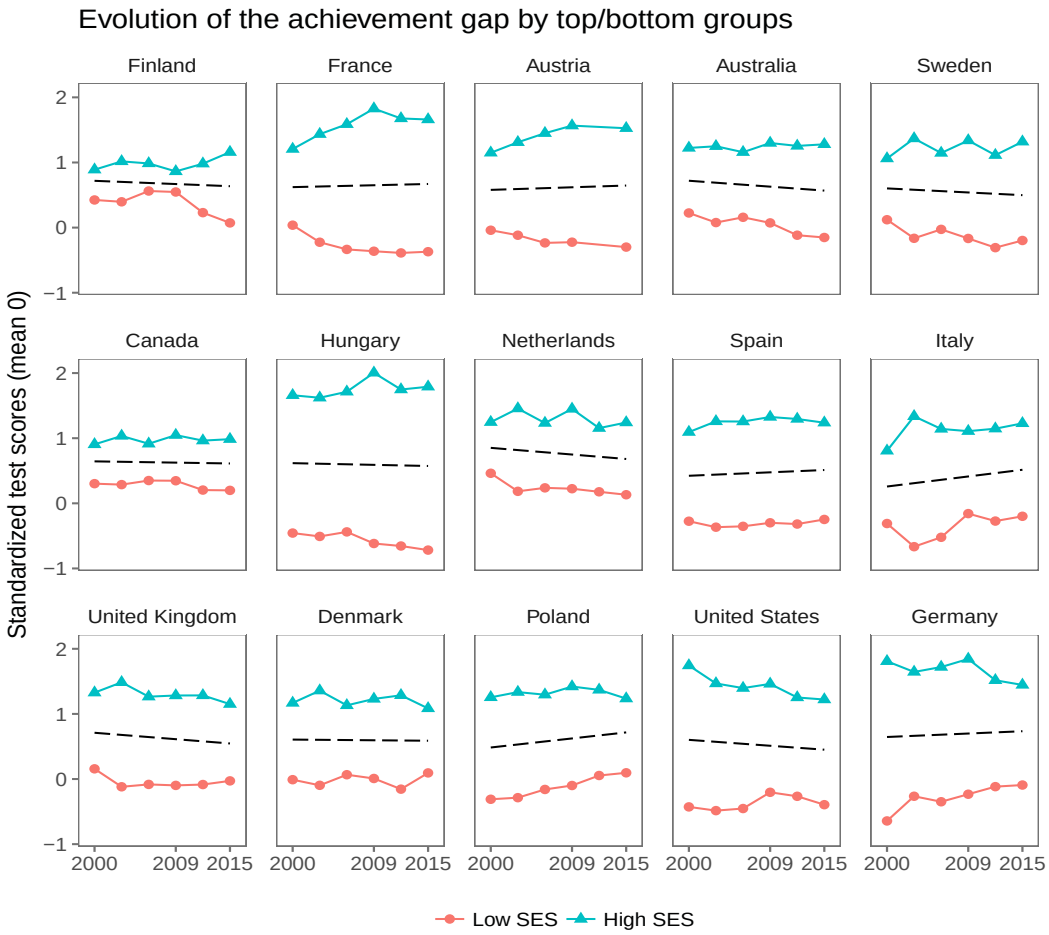

Figure 4: Evolution of the achievement gap for separate SES groups

Something reassuring is that mathematics and reading (not presented) follow basically the same trend across all countries [12]. This means that the result is not an artifact of chance. However, it's important to disentangle where is this gap originating from. Is this because the top are improving while the bottom decreases? Or is it that the bottom is catching up? Up next we plot the same graph but show the divergent patterns between high/low SES origins.

6.2 Source of the achievement gap

Figure 4 shows the evolution for both low and high SES groups separately. The middle line is a summary line showing whether one of the groups is growing/shrinking faster than the other. We find plenty of variation in these patterns. For example, in the United States the top seems to be equalizing much stronger than than the bottom is catching up. The U.K seems to be following the same path as the United States as well. On the other hand, in Poland the gap seems to be closing because the bottom SES group is catching up much faster. The Netherlands shows a similar pattern as the U.S, U.K patterns but upon closer inspection the explanation is very different. We see that the slope is negative (like in the previous two countries) but that's because the low SES group is decreasing at a greater rate than the high SES is decreasing.

We now pay attention to the flat summary lines. These lines should be interpreted with caution because it doesn't mean that the gap is not increasing. Denmark, Belgium, Netherlands, Norway and

[12] Reading had an even bigger decrease for similar countries, and a much smaller increase. The previos two graphs for reading are available upon request

Spain show a flat line because the gap is growing very little, if at all. On the other hand, Sweden, France, Finland, New Zealand, Austria, among other countries, show a flat line because both groups are distancing themselves at a very similar rate. These results highlight the importance of not only summarizing average achievement gaps. The source of these gaps vary greatly between countries and we see how each of these patterns contibutes to the overall inequality of a country.

After analyzing the trend for each country and where this trend is coming from we start to find some interesting results. We see that the United States is closing the achievement gap at a very rapid rate, but at the same time we notice that it's because the top SES group is going down faster than the bottom is catching up. In contrast, we see countries like Germany and Poland where the gap is closing as well but because bottom group is catching up faster than the high SES group is coming down. In fact, the Polish case is the most egalitarian of all the countries presented here; the bottom group is catching up at a rapid pace and the top group is maintaining their high levels of performance, fostering a trully equalizing effect.

Quite interestingly, in countries where the gap is widening, both groups are distancing themselves at a very similar pace. Finland, France and Austria, show this exact pattern. Give or take, we have a solid pattern that when the gap widens, it's because there's top-to-bottom inequality. But when the gap is being reduced, the source can vary between top and bottom changes.

To help assess these results even further, the source of this gap can be deconstructed further into two broader gaps. We want to know if these gaps have increased because the 90th SES group has distanced itself from the 50th SES group or because the middle has been also distancing itself from 10th SES group. For example, we're interested in searching which gap is bigger between countries, the 90th/50th or the 50th/10th? Moreover, are some of these gaps contributing more to the 90th/10th gap we saw earlier? Up next we plot all the combinations of the 90th/50th/10th gaps and pay attention to the magnitude of each gap in figure 5. Because we visualize three plots for every country, we only plot a selected number of countries that represent the overall patterns from the previous graphs [13].

We start off with Denmark. Generally speaking, we find that the 50th/10th gap is contributing more than the 90th/50th gap as the gap is bigger in mathematics. Furthermore, we find, just as in the 90th/10th SES gap, that there isn't a significant decrease [14]. France shows a similar pattern as Denmark in that the 50th/10th SES gap is the main contributer and that the three gaps have the similar trend.

Germany and U.S, on the other hand, show a very different landscape. For example, Germany, shows that the 90th/10th gap is primarily driven by the 50th/10th SES gap and that the slopes of the 90th/50th gap is much weaker than the other two gaps. This means that top/middle SES groups have shrunk much less than the other groups. The big catching up is actually coming from the middle/bottom SES groups. However, we don't know if this is because the bottom is catching up or that the middle is coming down. Based on the gap plots from above, we can infer that it is because the bottom is catching up rapidly because against the 90th SES group, it is increasing steeply every year.

The U.S shows a different pattern from all other countries. Both the 90th/50th/10th gaps seem to contribute rather similarly to the overall 90th/10th gap (although 50th/10th is marginally higher) but more interestingly, the 90th/50th gap has stagnated in the last 15 years. The top/middle groups have a relatively big gap and it hasn't changed at all. On the other hand, the steep decrease we saw for the U.S is coming exclusively from the 50th/10th gap shrinking. The reason, nonetheless, might be different from the German case, where it's probable that the bottom is catching up. Here, as in the 90th/10th gap, the middle might be 'dumbing' down closer to the low SES group. The big takeaway

[13] United States and Germany show a decrease, France shows a steep increase and Denmark shows no significant change.

[14] With the exception of reading in the 50th/10th SES gap which seemed to decrease from nearly 1.5 to 1 standard deviation.

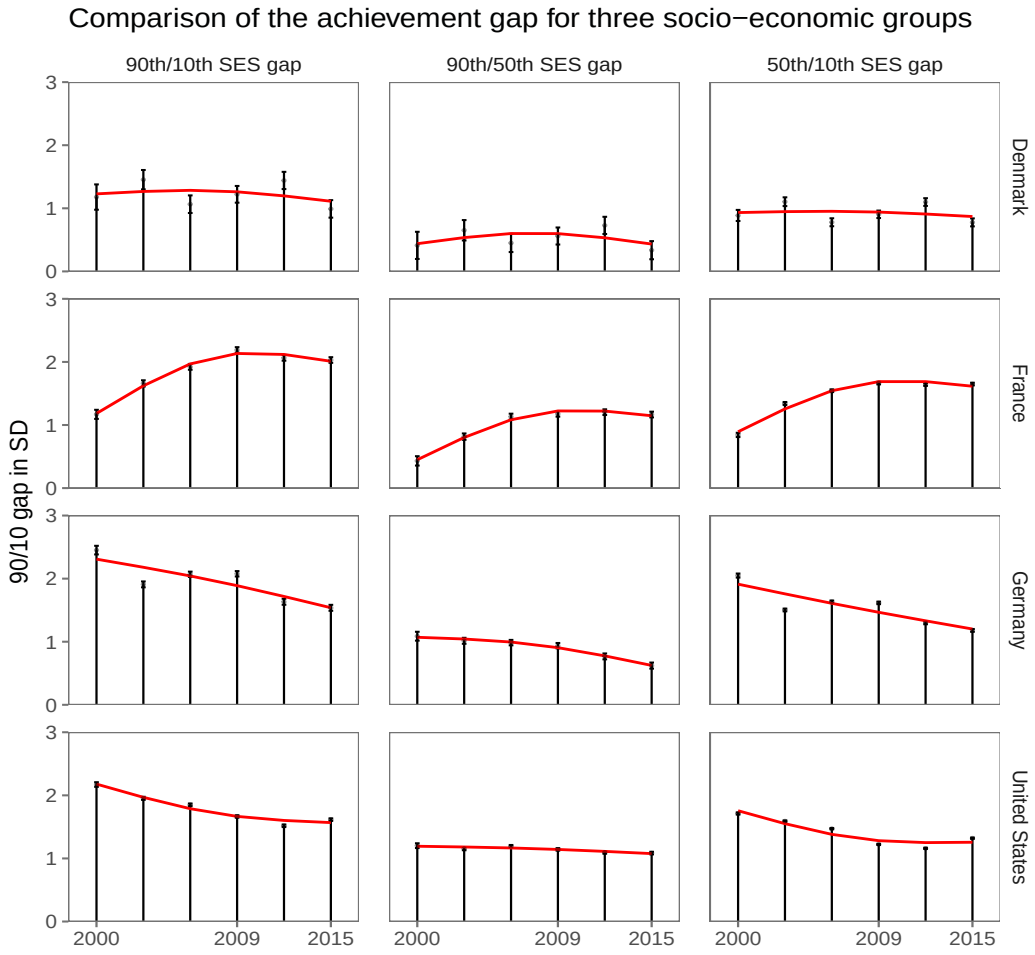

Figure 5: Evolution of 90/50/10 SES gaps for selected countries

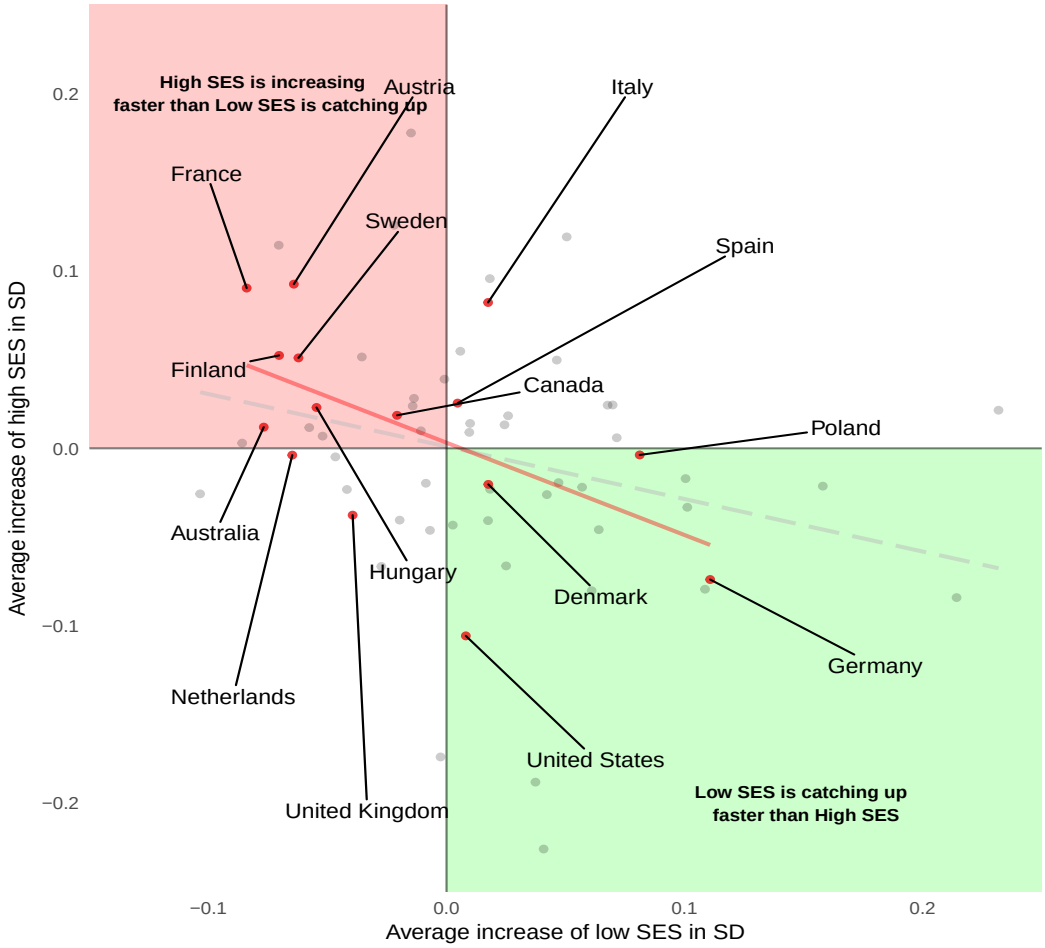

Figure 6: Rate at which top/bottom SES groups are catching up for the 90/10 SES gap

from the U.S is that the decrease of inequality is coming from the shrinking of the 50/10 gap.

In a more convenient way, we can inspect whether there is a lot of variance in the rate at which top/bottom groups are changing in figure 6.

We see that there are more countries where High SES is increasing inequality faster than low SES is catching up (top-left panel) but there's also a fair share where low SES is catching up faster, such as the United States and Germany. Regardless of this, we see an overall negative relationship where countries in which the low SES group is catching up, the high SES group goes down. This is very interesting because it suggests that there's a trend for countries to have the bottom SES groups are climbing upwards, a trait of meritocracy. However, this means that the top group comes down, rather than stagnate or increase. Equally important, we see very few countries where low SES are decreasing (so going even further down) and High SES is also decresing (bottom left panel). On the other hand, we do see more countries where the both groups are increasing (top right panel). We plot all remaining countries in the back, together with a trendline and we see that the relation holds for all PISA countries.

To sum up, we find no evidence in support of our first hypothesis. We uncover that the tracking setup of a country doesn't imminently mean a certain degree of equality/inequality. Countries where the top SES group is increasing and the bottom is decreasing (top-left panel) have dramatically different tracking setups. Austria is an early tracker with several tracks, while Sweden is a late tracker with only 1 track. On the opossite panel (bottom-right), where we find countries that the bottom group is catching up while the top is coming down, we find similar results. Denmark and Germany are opposite poles in terms of tracking, while the U.S is distinctive in that it has little formal curricular

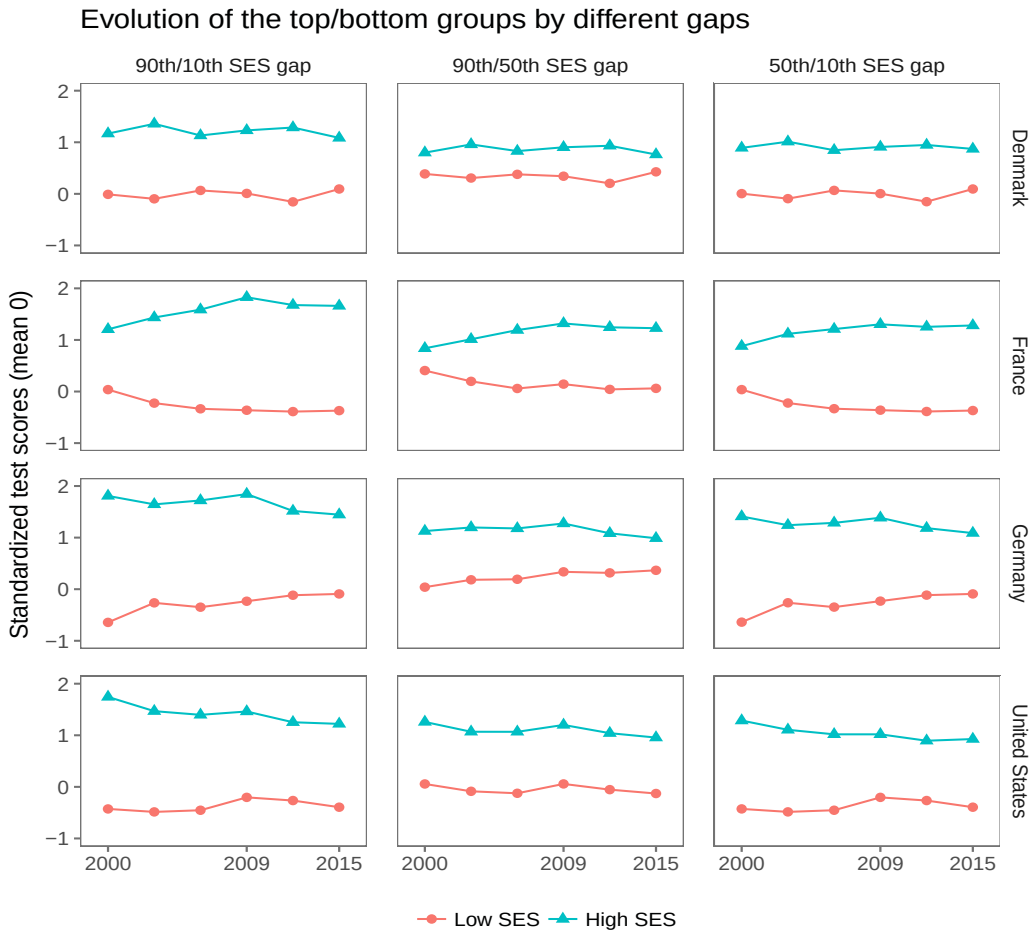

Figure 7: Evolution of the SES gaps for top/bottom groups

stratification. We see no definitive evidence of top-to-bottom equality in all of these countries (except for Poland). We find that regardless of the different tracking setups, we see that there's a tradeoff: either the bottom comes up and the top comes down or the top goes even further up and the bottom further down.

But, similar as before, we don't know which group is causing the decrease/increase of the gap in the 90th/50th and 50th/10th gap. For that we plot the evolution of these gaps in figure 7.

With Denmark there's no average change across all years. However, for France, we see that the gap is growing much faster for the 50th/10th SES gap. For Germany we see that the gap is closing in all three groups although much more strongly for the 50th/10th SES gap. It looks like the 50th/10th gap is mainly driving the changes in the gaps for most of the countries that actually see a change over time. For the U.S, there's very little change in the 90th/50th SES gap, with most of the change coming from the 50th/10th gap.

6.3 Equality-efficiency tradeoff

Moving on, we now discuss if there's a tradeoff between an educational system being efficient, i.e raising the overall performance of the country and whether that challenge compromises equality at the same time. For example, in order to improve the average score in, let's say, mathematics, tracking can improve the average score by making children in the higher trackers perform better, raising the overall performance. This, however, can come at a price. If tracking actually benefits the top but doesn't increase the performance of the bottom tracks then we see a net increase of inequality accompanied

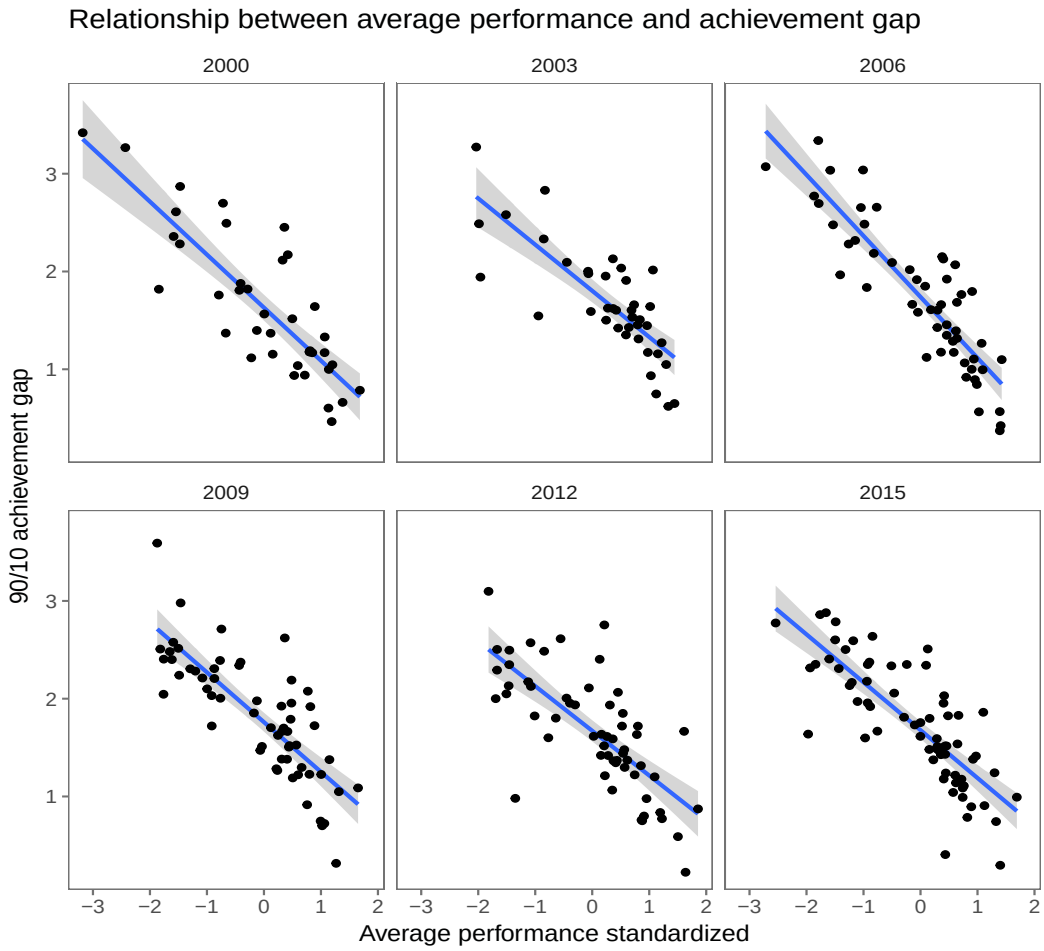

Figure 8: Relationship between 90/10 achievement gap and average performance in mathematics

with an increase in the overall performance. Given that we have the data for the average performance for each country and the achievement gap, we can investigate this relationship in detail. We first being by plotting all countries pooled for each year in figure 8.

For the past 15 years we see a very strong correlation suggesting that countries that perform very well in PISA have actually very low achievement gaps (or the other way around) [15]. The correlations are between -0.73 and -0.88. This confirms the theoretical and empirical frameworks suggested by Van de Werfhorst and Mijs (2010). However, what they didn't really go in depth was whether this corelation is strongest for some countries over time. However, we should not get trapped into an ecological fallacy and assume this relationship is identical within each country. With the data available we can look at the evolution separately for many countries. We plot it up next in figure 9.

Looking at figure 9, we see a strong correspondence between the two variables. In countries like France we see quite a functional form between both inequality and performance. As average performance plummets, the achievement gap is increasing. This pattern is also seen in Austria, Australia and Germany. Yet we also see the opposite trend. We see countries like Poland, Finland, Canada and less pronounced, Italy, where this pattern is evident. The surprising case here is the United States. We see that as the average performance has been decreasing, the average achievement gap has been decreasing as well. In fact, among the few countries we should find something that resembles this pattern should be the U.S because it doesn't have a strict institutionalized tracking system. However, the explanation behind this pattern is very speculative.

Finally, the last notable pattern is that in some countries the average performance is so high that

[15] We graph all countries-years pooled as well and the relationship is identical

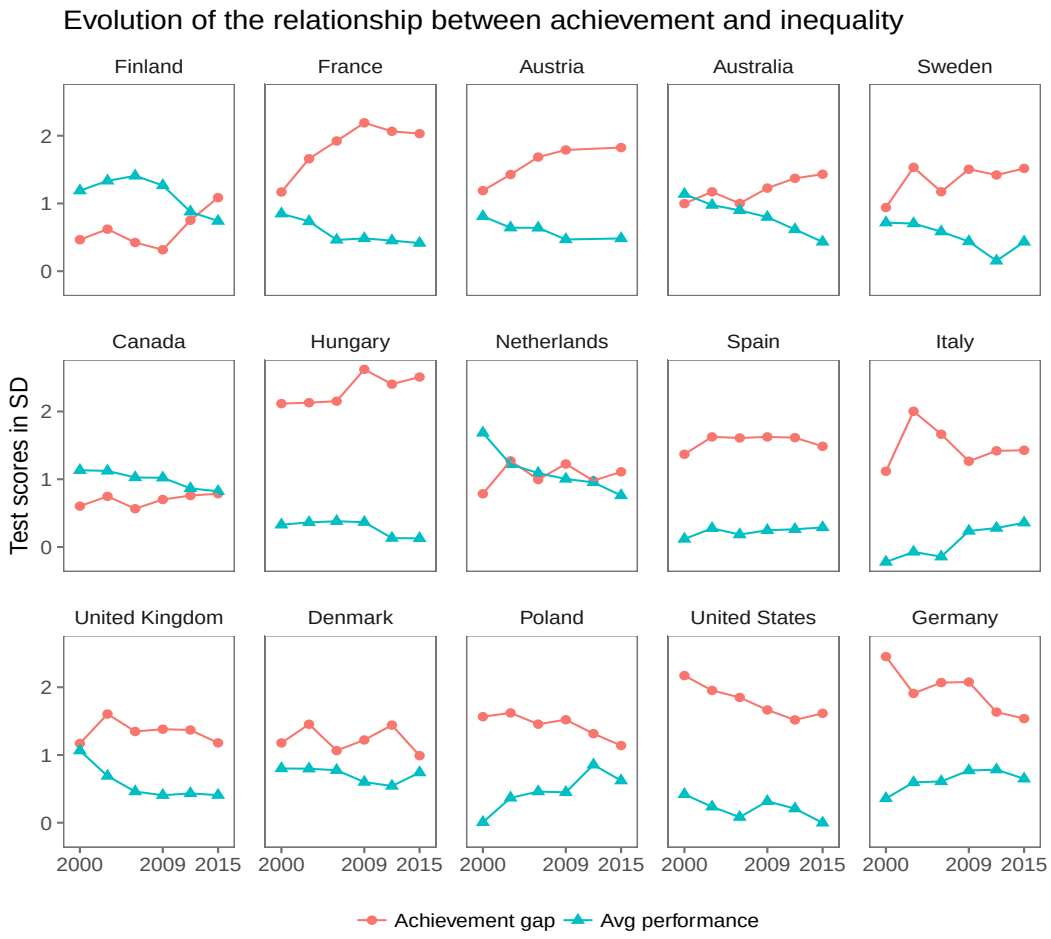

Figure 9: Time trends of achievement gap and average achievement between 2000 and 2015

it actually overperforms the achievement gap. Only Finland, Canada and Netherlands (three very equal countries) have their average performance in standard deviations above their level of inequality, which coincidentially, are countries that have the lower level of inequalities among the ones presented here.

These results actually confirm our second hypothesis, that there's a negative relationship between the two variables. We see countries like Poland, Canada and Germany where the pattern is just as the aggregate trend, namely that as the average performance increases, inequality seems to be reduced in parallel. Conversely, France, Austria and Australia, to name a few, portray the opposite: as average performance decreases inequality increases. But we also unconver important variation within countries in the relationship between average achievement and dispersion of achievement (have a look at the Netherlands, Italy and the U.S).

6.4 The role of curricular tracking in explaining inequalities

We now move to the final section of the paper, where we discuss what is a likely explanation to the patterns from before. The reason why countries differ in their evolution of inequality is still unknown. There is ample evidence showing the inequality between countries, to a certain extent, can be explained by the degree of tracking (Hanushek et al., 2006). But we are also interested in what explains the between country evolution of inequality, that is, the explanation as to why in certain countries it is increasing more than others over time.

This is clearly a difficult task given that, at least at the country-level, not many features of the educational system change in a 15-year time span, resulting in a practically invariant variable from which to explain the level of inequality (without any variance, we can't explain anything.). However, there are other strategies that can at least partially help us to understand the phenomena, and these are precisely the ones we follow.

Following the literature on the topic, we suspect that this evolution is in part very tightly related to the degree of stratification of the educational system. In a very simple frame-work, Hanushek et al. (2006) found that tracking indeed exarcebates inequality in the life course of children. Following the same strategy, we believe that tracking is also allowing the same gap in 15 year olds to grow over time, acompannied by other things. Yet, instead of concentrating only on the age of selection (just as they do), we explore a more fine grained definition of tracking, which is possible through the work of Bol and Van de Werfhorst (2013). Aside from the age at first selection into tracks, we also use the the number of tracks in the country, the percentage of the entire curriculum that is tracked and a vocational index [16].

Considering that our summarized data has very few observartions [17], we want to asses the uncertainty of our model by using a fully bayesian hierarchichal linear model. The benefits of this approach are twofold: first, we can asses the model fit much more intuitively, something very important given our sample size, and secondly, given that our countries are clustered into years, we can control for country heterogeneity within years, removing any bias in the standard errors.

In the final model, we regress the inequality of a country on a dummy variable where 1 equals only 1 track vs more than 1 track, whether the age of first selection is 15 or more vs below 15, a dummy stating whether the percentage of tracked curriculum is above 0 (that is, any tracking), and the standardized vocational index [18]. We are reluctant to include more variable in the model, first to keep it more parsimonious, and secondly to prevent overfitting due to our sample size. We present

[16] This index is a factor loading from a principal factor analysis of the percentage of students in upper secondary vocational education (taken from two sources, to reduce measurement error). We take all of the data from Bol and Van de Werfhorst (2013).

[17] We have per year, per country, one single inequality indicator, leaving the total sample size at 195

[18] We create dummies to avoid multicollinearity. The final model has a maximum VIF of 2.2.

the results in the table below.

	Model 1	Model 2	Model 3	Model 4
Only 1 track	-0.53 (-0.64/-0.41)	-0.31 (-0.44/-0.18)	-0.31 (-0.45/-0.17)	-0.21 (-0.34/-0.07)
Age selection >= 15		-0.38 (-0.51/-0.25)	-0.38 (-0.51/-0.26)	-0.57 (-0.71/-0.43)
% of curric tracked			0 (-0.23/0.23)	0.6 (0.3/0.89)
Vocational Index				-0.27 (-0.36/-0.19)
Intercept	1.77 (1.69/1.84)	1.94 (1.85/2.03)	1.94 (1.69/2.19)	1.54 (1.27/1.81)
R square:	23%	32%	32%	40%
Sample size:	195	195	195	183

Table 2: Explaining 90/10 achievement gap - Multilevel model

Given that the classical p values are inexistent in bayesian modelling, we present uncertainty intervals next to each coefficient. The first model only includes the 1 track dummy. This variable shows that countries that have one track, have on average, half a standard deviation less inequality than other countries. This covariate alone explains about 23% of the variance of the evolution of the 90/10 achievement gap. The second model now includes the dummy of whether the country has an age selection of 15 or more and we can see that countries with late selection into tracking have about -.38 less standard deviations in inequality. Also note that the number of tracks dummy was now reduced, suggesting these two variables are explaining similar things. This is expected given that late tracking is usually associated with fewer tracks, such as in the Scandinavian countries. The third model now includes whether the country has any degree of tracking, meaning whether the country has above 0% of the curriculum tracked. The coding of the variable comes from visual inspection of the distribution, where some countries, such as the United States and Canada, have untracked curriculums. For this last variable we see no relationship at all, but after including the vocational index, the model is completely reversed with countries with tracked curriculum having about 0.6 more standard deviations in inequality that countries with no tracking.

We also note that vocational enrollment is associated with less inequality. Not surprisingly, the age of selection coefficient, also increased due to the correlation between age of selection and vocational enrollment. Countries with high vocational enrollment are also those with early age selection, such as Germany and Austria. These two variables should be inspected further, perhaps with an interaction, given that both coefficients are sizable and very correlated. This final model explains over one third of the variance in the evolution of the 90/10 SES gap, with an R square of 40%. The uncertainty of each of the covariates varies a lot with the percentage of curriculum tracked dummy ranging all the way from 0.9 to 0.3. Nonetheless, we see that the other coefficients are very strong and have reasonable intervals. It is very important to consider the uncertainty of our estimates, to gain confidence in our model fit. So far, this model seems to fit the data quite reasonably.

All of these countries have a different tracking structure, with some setups being more egalitarian than others. Using the previous model we make a simulation and predict the level of the 90/10 achievement gap for all possible tracking setups for all countries. Below we plot all these combinations. The X axis has the number of tracks that a country has, the Y axis has the predicted 90/10 achievement gap and below each country name we have whether the country has an age selection of 15 or more or below 15. In addition, to confirm how accurate the model is, we also plot the actual level of inequality in the actual tracking setup of the country. Let's take Germany for example.

The star point for Germany is in the '>1 track' and '<15 age' setup, meaning that Germany has a tracking system that has more than 1 track and an age of selection of below 15. We see that the model predicts the level of inequality for that set up quite accurately (height of the bar), relative to the actual level (the star point). But if the setup would be in the ideal '1 track' and age of selection above or equal to 15, the inequality would be much lower.

These graph reveals some interesting patterns. Generally speaking, we find that the simulation

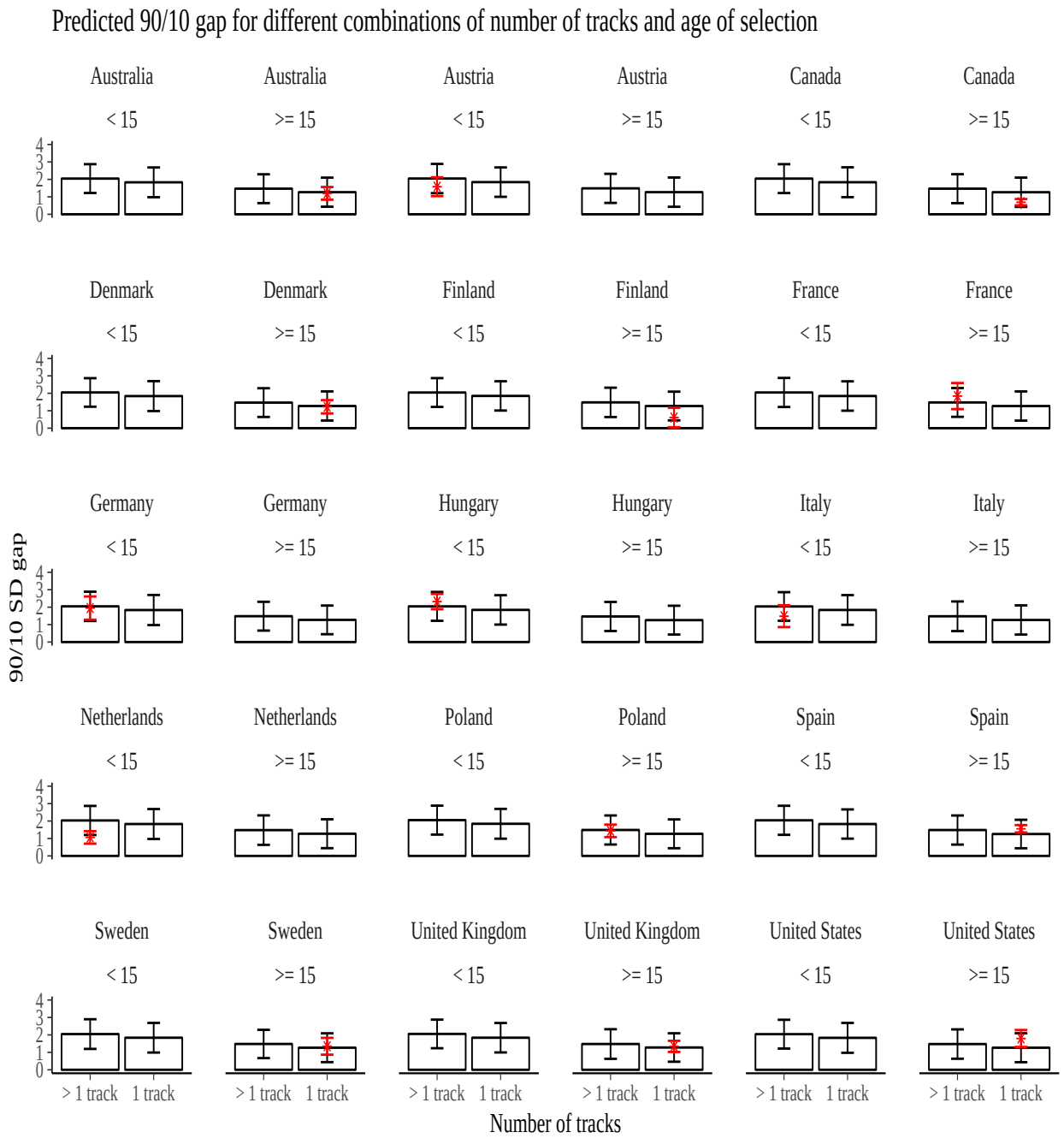

Figure 10: Simulation: predicted 90/10 achievement gap for different tracking setups

predicts that virtually all countries which are not in this 'ideal' tracking, that would switch to the 'ideal' setup, would experience a reduction of the achievement gap. Yet we also find the opposite result. Countries which are in the 'ideal' setup that would change to the 'worst' setup would see a widening of the achievement gap. On average, we find that if all countries switched to the 'ideal' setup, we would see an average reduction of 20%. This is a very important figure considering that Reardon (2011) found that the US gap increased by about 40-50% in 30 years. Had all countries in the 'ideal' tracking switched to the 'worst' tracking, we would see an average widening of nearly 72%. For those countries which are not in the 'ideal' tracking setup, we compute below the percentage change from switching to the ideal tracking.

Country	Number of tracks	Age of selection	Inequality	Predicted	% reduction
Austria	More than one track	Less than 15	1.58	1.27	20%
France	More than one track	15 or more	1.84	1.27	31%
Germany	More than one track	Less than 15	1.95	1.27	35%
Hungary	More than one track	Less than 15	2.32	1.26	46%
Italy	More than one track	Less than 15	1.48	1.26	15%
Netherlands	More than one track	Less than 15	1.06	1.27	-19%
Poland	More than one track	15 or more	1.44	1.27	12%

Table 3: Simulation: reduction of achievement gap if countries switched to 'ideal' tracking

The biggest reductions are for both Hungary and Germany, with 46% and 36% respectively. France, which had a particularly steep increase in the last 15 years, would see a reduction of the 90/10 SES gap of 31% and Austria, with a very similar tracking structure as the German case, would see their achievement gap reduced by 20%. The very suprising result is the Dutch case, for which a switch from the 'worst' tracking setup to the 'best' case would result in an increase of the gap of 19%.

This last result could mean two things. First and foremost, the Dutch educational system has built other mechanisms that, despite the tracking system, help children from low social origins escalate upwards. This explanation is not very trivial considering that in most mobility research, the Netherlands stands out as a very equal country relative to other countries with a similary tracking structure, like Germany (Shavit and Blossfeld, 1993; Esping-Andersen and Wagner, 2012). Another possible explanation, and something which needs to assessed, is whether our model is predicting accurately. If we look at the plot from above, we see that for some countries, the model is a bit off from the actual level of inequality. Moreover, although we see that if all countries switched to the 'ideal' tracking setup we would see a vast reduction in inequality, we also see that each of these predictions is followed by a wide uncertainty interval.

Considering that the results are very consistent for every country, and in most countries the prediction matches the actual value, this could simply be a consequence of low sample size. Whether these results stand the test of a different empirical strategy or a higher sample size, we leave up to future research.

A more summarized empirical strategy is to actually standardize all the tracking indicator into one index, something that Bol and Van de Werfhorst (2013) already do quite nicely. Using this standardized tracking variable and the same bayesian multilevel model, we now model the tracking and vocational index as before, but add an interaction between two, given that in the previous model we recorded a strong shift in the model when tracking-related variables and the vocational model where both included in the same model. Below are the results for this model.

We confirm the results from the previous models, where more tracking is associated with an achievement gap of about 0.4 standard deviations wider and more vocational enrollment is associated with a reduction of about 0.2 standard deviations, half the slope of tracking. Meaning that the tracking effect might be offsetting the effect of the vocational index given that it's half the size of the slope. Both these coefficients are actually significantly different from each other, so it is possible. Before we

	Model 1	Model 2	Model 3
Tracking Index	0.26 (0.21/0.32)	0.33 (0.27/0.4)	0.41 (0.34/0.48)
Vocational Index		-0.19 (-0.26/-0.11)	-0.22 (-0.29/-0.15)
Trackin * Vocational Index			-0.12 (-0.17/-0.06)
Intercept	1.6 (1.53/1.67)	1.65 (1.58/1.72)	1.71 (1.63/1.78)
R square:	24%	30%	34%
Sample size:	195	183	183

Table 4: What explains 90/10 achievement gap? Tracking and Vocational interaction

explore it visually, note that the interaction between the two is very reliable and significant, considering that small sample size. Below we graph the interaction for different quantiles of the vocational index.

The results are very interesting. We see that for lower levels of tracking (meaning little to no tracking) theres a very low achievement gap, regardless of whether the country has very high or very low vocational enrollment. However, once tracking enters the picture, vocational enrollment can be a strong equalizer. We see that the bottom quantile of vocational enrollment has almost twice the achievement gap as the top quantile for vocational enrollment. Moreover, our earlier interpretation (that tracking is more important than vocational enrollment for reducing the achievement gap) still holds and is very evident here. Despite the equalizing power of vocational enrollment, a high level of tracking with the highest level of vocational enrollment still leaves a country with an achievement gap of over 1.5 standard deviations. However, a country with low levels of tracking and *any* level of vocational enrollment is nearly 1.3 standard deviations at best. A key finding from these results is that tracking is a reasonable explanation of the 90/10 achievement gap, and we should consider it as a very important component in the reduction of the achievement gap. However, these results highlight that it should not be studied in isolation from other features such as the degree of vocational enrollment. We see a very sizable change in the achievement gap for a given country with the same level of tracking and high and low levels of vocational enrollment.

The exact mechanisms through whick the tracking setup is explaining these inequalities is still very speculative. We could argue that tracking might've introduced the inequality once it was implemented. Another possibility is that tracking, althought not changing, is currently exarcebating it every year. Another explanation, which we find strong evidence for, is that the tracking setup interacts very closely with the vocational setup. These explanations are very speculative and further research should carry the task of looking for a more concrete explanation.

To test whether these results hold for other samples, we compute the same models as before but also for the 80/20 and 70/30 SES gap. Up next we plot these coefficients for the three SES gaps.

With the exception of the Intercept, the coefficients of the three models look pretty much the same. The tracking index seems slightly stronger for the 90/10 model but the uncertainty intervals overlap with the other two models. These results give our claims much more credibility.

Finally, until now, we've studied how we can explain the evolution of the 90/10 achievement gap, by clustering on the year dimension of each country and then performing a linear multilevel model. This approach was adopted given that tracking hardly changes over time, making it impossible to model the change over time directly. However, there is another way of estimating the evolution of the achievement gap, which we try up next.

Instead of estimating the evolution over time, we calculate the cumulative change over time for each country, summarizing the 6 year-time span of achievement gaps into one cumulative inequality gap. This approach dramatically reduces the total sample size to 34, the number of total countries available [19]. It is for this reason that we only include at most 2 covariates in the model, the tracking

[19]We've been presenting a selected number of countries given the lack of space, but for every model/estimation we

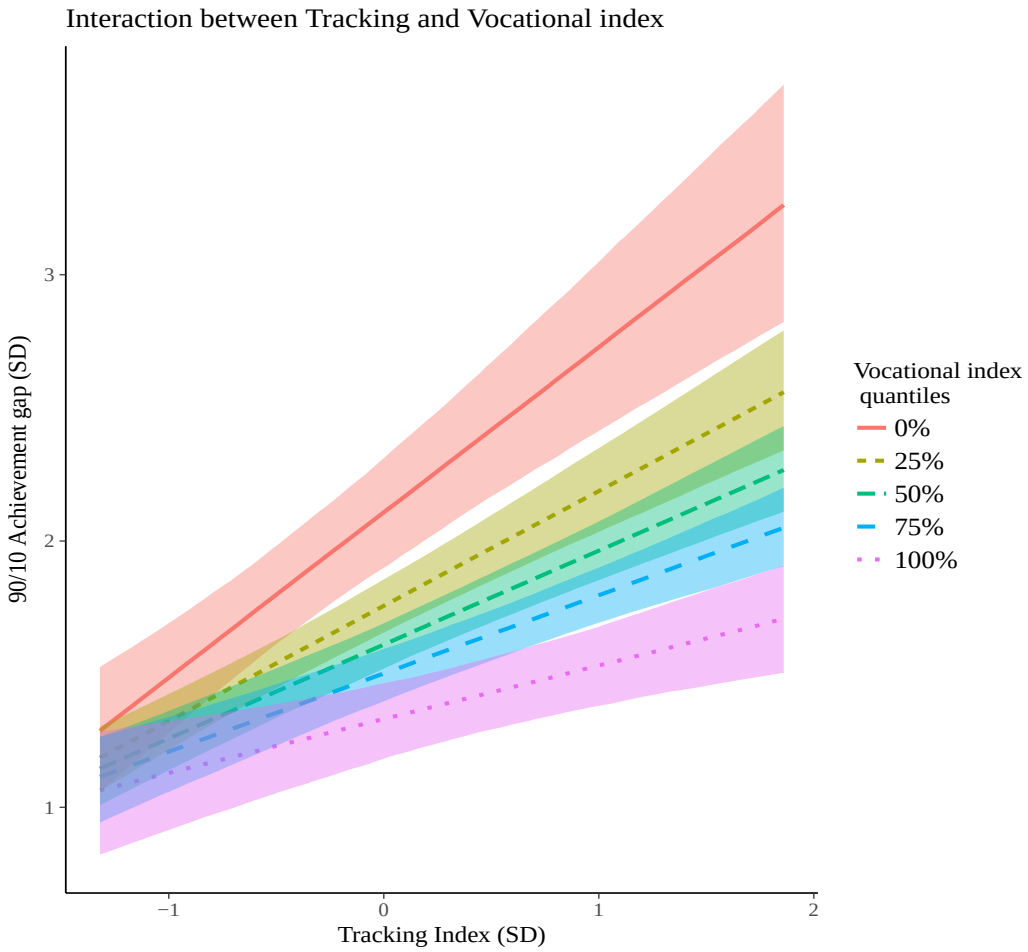

Figure 11: Interaction between tracking and vocational values. Legend ordered relative to the lines

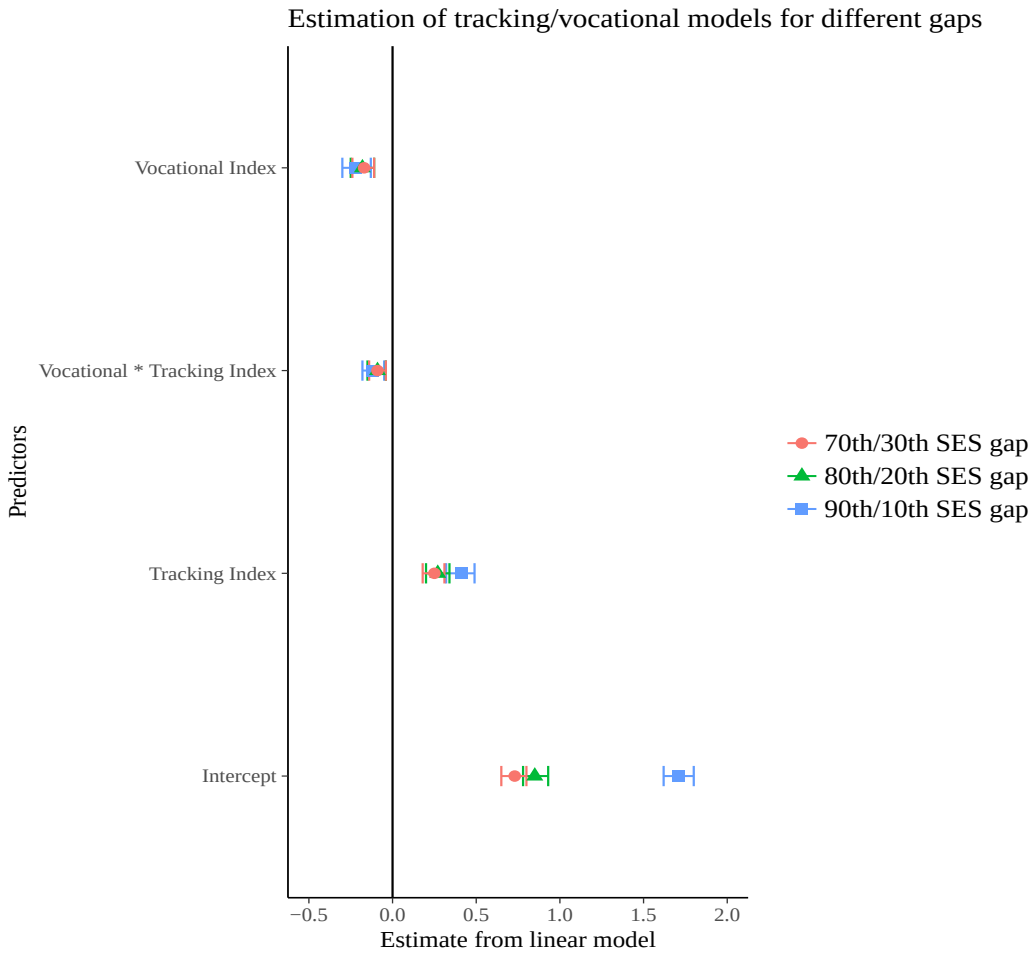

Figure 12: What explains achievement gaps? Model comparisons

and vocational index, which both summarizes the previous models and keeps it parsimonious.

	90th/10th	80th/20th	70th/30th
Track Index	1.66 (0.85/2.49)	1.11 (0.46/1.78)	1 (0.37/1.67)
Vocational Index	-1.04 (-1.95/-0.12)	-0.9 (-1.64/-0.14)	-0.87 (-1.6/-0.16)
Intercept	9.37 (8.6/10.12)	4.6 (3.99/5.19)	3.89 (3.3/4.47)
R square:	30%	25%	23%
Sample size:	32	32	32

Table 5: What explains the cumulative achievement gaps? Models for three different achievement gaps

We see that 1 standard deviation increase in the tracking index widens the 90/10 SES gap by 1.66 standard deviations over time, and the vocational index reduces it by 1 standard deviation and these two covariates explain nearly one third of the cumulative gap. As we decrease the achievement gap (80/20, 70/30) these coefficients become smaller, but keep their strength and their predictive power. Comparing this model to the previous models, we see a very similar landscape, but an even more reassuring one, given than we're tackling the evolution of the achievement gap directly with this indicator. We can safely say that these results, at worse, are supporting the previous findings and making the claim even more credible.

References

Bauchmüller, R., Gørtz, M., and Rasmussen, A. W. (2014). Long-run benefits from universal high-quality preschooling. *Early Childhood Research Quarterly*, 29(4):457–470.

Bernardi, F. and Ballarino, G. (2016). *Education, occupation and social origin: a comparative analysis of the transmission of socio-economic inequalities*. Edward Elgar Publishing.

Björklund, A. and Jäntti, M. (2009). Intergenerational income mobility and the role of family background. *Oxford Handbook of Economic Inequality, Oxford University Press, Oxford*, pages 491–521.

Bol, T. and Van de Werfhorst, H. G. (2013). Educational systems and the trade-off between labor market allocation and equality of educational opportunity. *Comparative Education Review*, 57(2):285–308.

Bradbury, B., Corak, M., Waldfogel, J., and Washbrook, E. (2015). *Too many children left behind: The US achievement gap in comparative perspective*. Russell Sage Foundation.

Breen, R. and Goldthorpe, J. H. (1997). Explaining educational differentials: Towards a formal rational action theory. *Rationality and society*, 9(3):275–305.

Breen, R. and Jonsson, J. O. (2007). Explaining change in social fluidity: Educational equalization and educational expansion in twentieth-century sweden 1. *American Journal of Sociology*, 112(6):1775–1810.

Byun, S.-y. and Kim, K.-k. (2010). Educational inequality in south korea: The widening socioeconomic gap in student achievement. In *Globalization, changing demographics, and educational challenges in East Asia*, pages 155–182. Emerald Group Publishing Limited.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., and Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the abecedarian project. *Applied developmental science*, 6(1):42–57.

include all countries available.

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R., and Narang, J. (2016). The fading american dream: Trends in absolute income mobility since 1940. Technical report, National Bureau of Economic Research.

Chmielewski, A. K. (2016). Changes in socioeconomic achievement gaps in international comparison, 1964-2012. In *2016 APPAM International Conference*. Appam.

Chmielewski, A. K. and Reardon, S. F. (2016). Patterns of cross-national variation in the association between income and academic achievement. *AERA Open*, 2(3):2332858416649593.

Coleman, J. S. et al. (1966). Equality of educational opportunity.

Cunha, F., Heckman, J. J., Lochner, L., and Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, 1:697–812.

Duncan, G. J. and Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. *Whither opportunity*, pages 47–70.

Duyme, M., Dumaret, A.-C., and Tomkiewicz, S. (1999). How can we boost iqs of dull children?: A late adoption study. *Proceedings of the National Academy of Sciences*, 96(15):8790–8794.

Erikson, R., Goldthorpe, J. H., and Portocarero, L. (1979). Intergenerational class mobility in three western european societies: England, france and sweden. *The British Journal of Sociology*, 30(4):415–441.

Esping-Andersen, G., Garfinkel, I., Han, W.-J., Magnuson, K., Wagner, S., and Waldfogel, J. (2012). Child care and school performance in denmark and the united states. *Children and Youth Services Review*, 34(3):576–589.

Esping-Andersen, G. and Wagner, S. (2012). Asymmetries in the opportunity structure. intergenerational mobility trends in europe. *Research in Social Stratification and Mobility*, 30(4):473 – 487. Consequences of Economic Inequality.

Fryer, J, R. G. and Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, 86(2):447–464.

Gamoran, A. (2001). American schooling and educational inequality: A forecast for the 21st century. *Sociology of education*, pages 135–153.

Ganzeboom, H. B. (2010). A new international socio-economic index (isei) of occupational status for the international standard classification of occupation 2008 (isco-08) constructed with data from the issp 2002-2007. In *Annual Conference of International Social Survey Programme, Lisbon*, volume 1.

Ganzeboom, H. B. and Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social science research*, 25(3):201–239.

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

Hanushek, E. A. and Wößmann, L. (2007). The role of education quality for economic growth.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902.

Jæger, M. M. and Holm, A. (2007). Does parents economic, cultural, and social capital explain the social class effect on educational attainment in the scandinavian mobility regime? *Social Science Research*, 36(2):719–744.

Magnuson, K. and Waldfogel, J. (2008). *Steady gains and stalled progress: Inequality and the Black-White test score gap*. Russell Sage Foundation.

Micklewright, J. and Schnepf, S. (2007). *Inequality and Poverty Re-Examined*. Oxford Univ. Press.

OECD (2002). *Education At A Glance: OECD Indicators 2002*. OECD Publishing.

OECD (2012). *PISA 2009 Technical Report*. OECD Publishing.

OECD (2016). *PISA 2015 Results (Volume I)*. OECD Publishing.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither opportunity*, pages 91–116.

Reardon, S. F. and Portilla, X. A. (2015). Recent trends in socioeconomic and racial school readiness gaps at kindergarten entry. *Stanford, CA: Center for Education Policy Analysis. Retrieved December*, 20:2015.

Saw, G. K. (2016). Patterns and trends in achievement gaps in malaysian secondary schools (1999–2011): gender, ethnicity, and socioeconomic status. *Educational Research for Policy and Practice*, 15(1):41–54.

Shavit, Y. and Blossfeld, H.-P. (1993). *Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Social Inequality Series*. ERIC.

Van de Werfhorst, H. G. and Mijs, J. J. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual review of sociology*, 36:407–428.

Waldfogel, J. (2006). What do children need? *Juncture*, 13(1):26–34.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31(2-3):114–128.

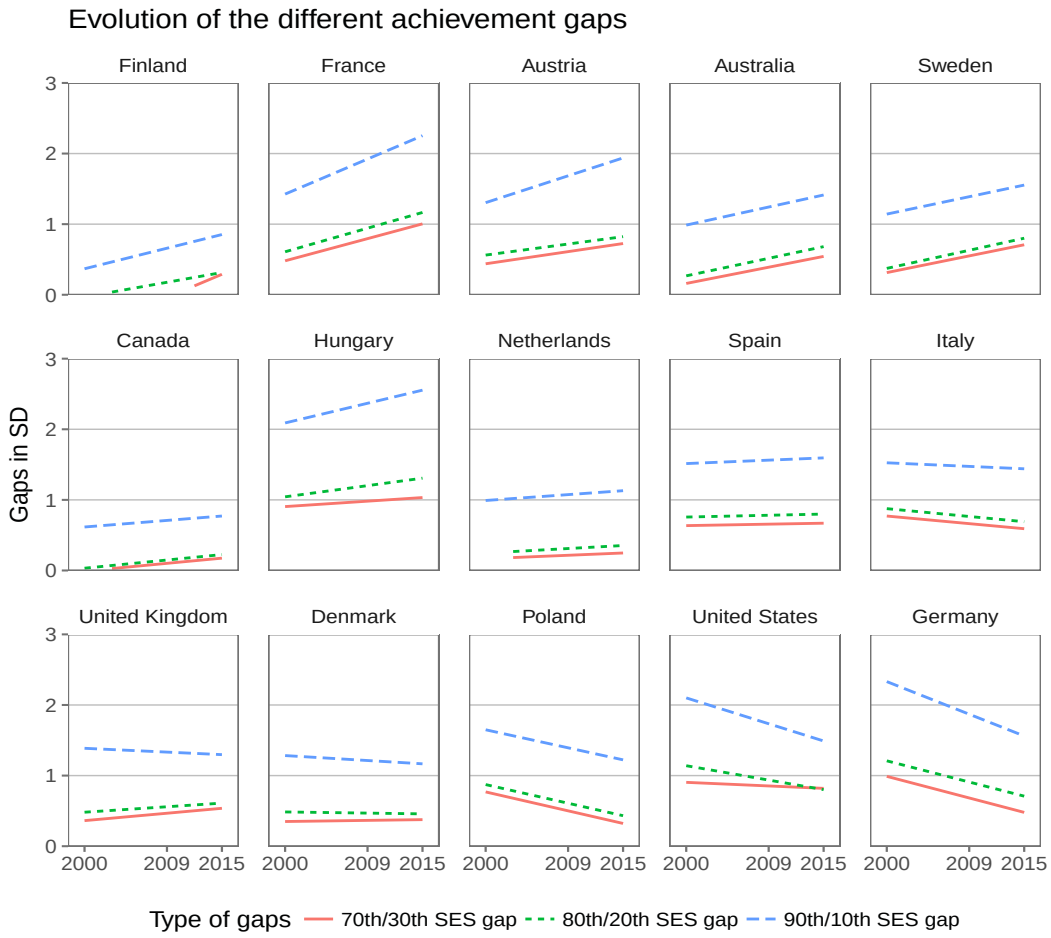

Figure 13: Evolution of the achievement gap for several gaps

7 Appendix

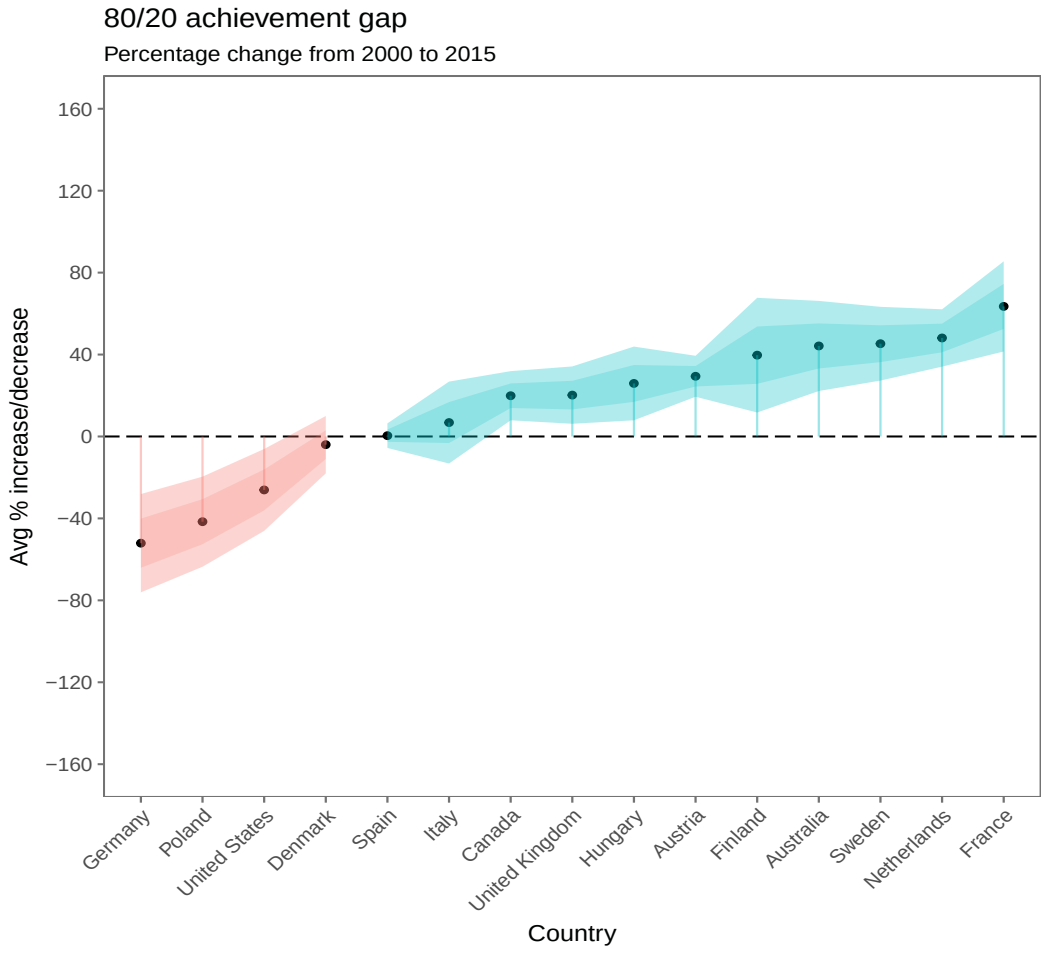

Figure 14: Percentage change in the 80/20 achievement gap from 2000 to 2015

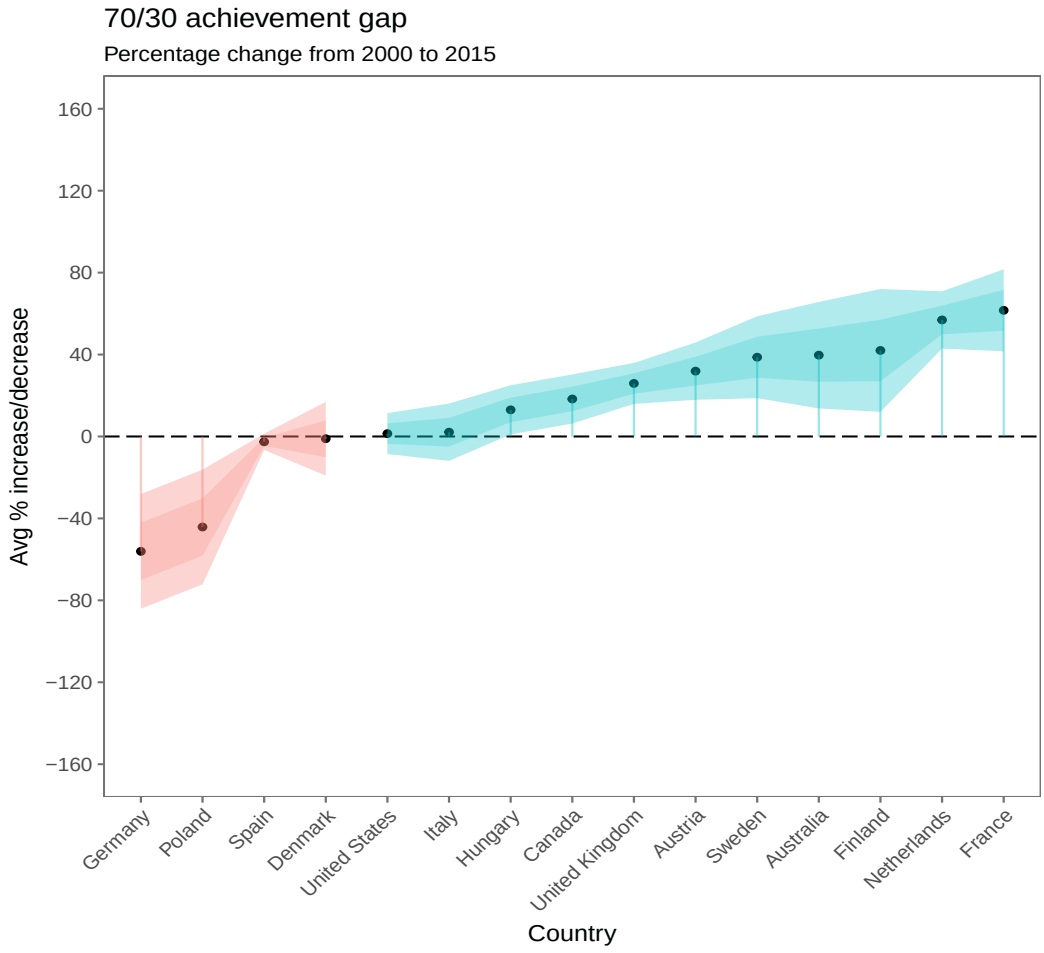

Figure 15: Percentage change in the 70/30 achievement gap from 2000 to 2015