Trade-offs between Equity and Excellence in Academic Performance: Evidence from 27 Countries.

Abstract

Educational policy makers traditionally perceived there to be a trade-off between educational excellence and educational equity. With the rise of cross-national comparative datasets, however, research has begun to suggest such a trade-off does not exist and indeed, higher variance in achievement and more segregated school systems may be associated with lower performance. However, such research has tended to focus on between nation analysis for which important covariates are not controlled (e.g., response set differences, latent cultural differences, etc.). Likewise, relatively little consideration has been given to whether a trade-off may exist for high or low performing students. Using five cycles of the PISA database, the current research explores within country trajectories in achievement and inequality measures to test the hypothesis of an excellence/equity trade-off in academic performance. Rejecting the trade-off hypothesis, we find a robust negative relationship between performance and inequality which is of statistical and practical significance. Detailed analysis of countries with large changes in average achievement from 2000 to 2012 suggest a focus on low and average performers may be critical to successful policy interventions within a given country.

Introduction

A critical issue in education relates to balancing concerns about maximising educational outcomes with ensuring equity both in terms of equal opportunity and in minimising excessive variation in those outcomes. This has been an ongoing concern in relation to social mobility research (Burger, 2016), educational attainment (Goldthorpe, 2007), and to a lessor degree concerns about performance in standardized tests (Checchi, 2006). Our paper is primarily concerned with issues relating to the association between standardized test performance (educational excellence) and the degree of variation in performance within a nation (our measure of educational equity). It is our contention that greater inequality in the variance of test scores will be negatively associated with average educational achievement. In this way we seek to directly challenge views that countries educational policies must make implicit trade-offs between educational excellence and equality. To test this hypothesis we consider a range of variance or inequality measures. Unlike previous research we focus on a) changes that occur within countries over time; and b) on where the changes occur in the academic achievement distribution. In the following sections we first position our research within the broader domain of educational equity research before outlining competing positions on the excellent/equity trade-off in educational ability. Finally, we consider what empirical research currently suggests about this debate and the limitation with the existing evidence base that we seek to overcome.

Educational Equity Research

Research concerning issues of educational equity spans much of the social sciences including economics (Checchi, 2006), educational policy (Rowe & Lubienski, 2017), sociology (Jerrim et al., 2016), and psychology (e.g., Parker et al., 2017). Given such broad concern, it is perhaps not surprising that this domain space includes a wide-ranging spectrum of concern and a considerable degree of variability in relation to the mechanisms addressed. This includes research and theory on intergenerational social mobility and transmission of educational attainment (e.g., Burger, 2016), and social class, ethnicity, and gender inequality in educational attainment and opportunities (see Goldthorpe, 2007 for a review). Not only do these research projects differ in outcomes of interest (e.g., adult income, years of education obtained, type of education obtained), but also in what purpose they perceive education playing in the socio-political and economic context (e.g., as human capital development, as a mechanism for distributing places in society, as providing opportunities for advancement, or for the cultivation of citizens) and in the methodology they pursue (e.g., a focus on correlations between parent and child years of education or the relative odds of an individual from a given educational origin transitioning into a given educational destination; Goldthorpe, 2007). As such it is critical that we are clear in the definition of the factors we consider here and our purpose.

Excellence. We consider excellence in relation to average ability scores at age 15 in international standardized tests. This is somewhat akin to Pfeffer's (2015, p. 353) definition of quality as evidence of "capabilities that serve as the functional prerequisite for social integration". Note that Pfeffer's definition implies an aim of educating students at least up to the point that they reach the threshold required for social integration. This reflects a socio-political focus that education should develop citizens (see Walzer, 1983) and that this need is of increasing significance and urgency in a world of growing complexity, interconnectedness, and plurality (Nussbaum, 1998). Its difficulty comes in defining where this threshold is, whether it differs across nations, and subsequently how quality should be defined in an international context. It is worth noting for example that Pfeffer does not define a threshold or set of thresholds, but rather considers quality as continuous measure of achievement in international adult assessments.

In contrast, we explicitly embrace excellence as continuous both in measurement and in definition. This approach is promoted by the OECD, who state that quantitative improvements in standardized achievement scores in international tests has an association with increased human capital and productivity as measured by economic instruments such as Gross Domestic Product (Hanushek, & Woessmann, 2010). This is not to deny the other roles that education plays, rather, this is an initial step that is broadly inclusive of social science foci, does not require the consideration of thresholds, but still holds policy implications.

Equity. The vast majority of sociological research on equity in education has focused on the relative chances of individuals within a given social class, ethnicity, or gender in obtaining a given number of years of education, general educational continuation, or educational pathways at points of differentiation (Lucas, 2001). Thus for example, Maximally Maintained Inequality (Raftery, & Hout, 1993), Effectively Maintained Inequality (Lucas, 2001), Rational Action Theory (Breen & Goldthorpe, 1997), and Modernization Theory (Marks, 2013), focus predominately on attainment of years of education, credentials, or long-term status attainment. Academic ability is generally used as a critical control or causal mechanism in these theories, where for example research considers differences in educational attainment for equally able children from different social groups.

Our focus, however, is more consistent with studies focusing on distributional concerns in academic ability (e.g., Hung, 2009). It is within these definitions that we consider whether increased excellence must come at the expense of expanded variance as has frequently been implied by educational theory and policy (see Pfeffer, 2015;Van de Werfhorst & Mijs, 2010 for a review).

Excellence versus Equality

Debate over excellence in education often suggests that educational systems produce the highest average performance if schools can tailor offerings to different levels of the underlying talent distribution of the student population (see Hoxby, 2003; Van de Werfhorst & Mijs, 2010 for a review; see Walberg, 2000 for an applied introduction). Checchi (2006, see also Hoxby, 2003) provides a detailed treatment of this argument but, put simply, it stipulates that, in the absence of government interference, families will choose a level and type of education for their children that will maximise the child’s achievement and, should this occur for most children, maximise the achievement of the nation as a whole (Friedman, 2002; Hoxby, 2003). At the core of this idea is that differentiated, stratified, decentralized, and/or private or privatised education (see Bol et al., 2013; Kerckhoff, 1995; Parker et al., 2016 for a review) provides a context that prepares children with different underlying talent with appropriate skills. This may mean less talented children are provided with educational content specifically focused on vocational skills (see Brunello & Checchi, 2007 for an overview). For talented children, no longer hampered by the need for teachers to limit the scope and speed of content for the benefit of less talented children, increased education system variance will maximise their learning gains (see Van de Werfhorst & Mijs, 2010).

Under this model, increased academic *excellence* for a country will tend to be associated with greater variance in achievement than *equal* systems due to selection effects, signalling, and different educational content (Jakubowski, Patrinos, Porta, & Wisniewski, 2010; Parker et al., 2016; Pfeffer, 2015). Thus there is a potential conflict in policy between maximising excellence (maximising average levels of achievement by allowing children to match their education to their potential) and ensuring there is equity (minimising the variability in outcomes between children; Gans & King, 2014). According to the trade-off position, excellence (i.e., high performing) comes at the cost of variability in results. But does the empirical evidence support this?

Excellence/Equality trade-off. Underlying the excellence/equity trade-off concern is the central tenet that an excellent educational system is, by necessity, counter to a system with limited variability. State policymakers will thus need to balance the trade-off between these competing goals. However, this trade-off is thought to have several parts. First educational differentiation or school choice means that different children receive different levels or types of education. Second, it may be that increased variance occurs due to mechanisms unrelated to government policy or, at least, unrelated to government *education* policy. For example, increased variance may come about due to wider social stratification be it by race, ethnicity, or social class (e.g., Rowe & Lubienski, 2017). Third, there may be barriers that prevent children from disadvantaged backgrounds gaining access to the type of education best suited to their underlying talent. Indeed, due to access to economic or other resources, risk adversity, or poor decision-making, parents might choose a type of education that is inappropriate for the child and thus require policies that provide such children with educational chances more in keeping with their ability (Friedman, 2002; Gans & King, 2014). Often suggestions for policy interventions indicate that, apart from ensuring talented children are not misplaced, no checks should be placed on the variance in academic ability within a country. Under such a system, achievement differentiation, decentralization, privatization, and stratification should be encouraged as they increase the options available to parents and improve overall performance. Government intervention, on the other hand, should focus only on reducing risks of student misplacement within this system (Friedman & Friedman, 1980).

Empirical Evidence. The underlying theory of the trade-off between excellence and equity argument is elegant. Yet it has been increasingly scrutinized by empirical evidence derived largely from studies using large-scale international student assessments (see Van de Werfhorst & Mijs, 2010 for a review). Anecdotally, criticism of the implied excellence/equity trade-off comes from the observation that high performing countries like Finland appear to combine both high levels of equality (including both low barriers to entry and relatively undifferentiated education) with high levels of performance in international tests (Simola, 2005). Empirically, evidence questioning the excellence/equity trade-off comes from two sources: a) empirical results that suggests the major sources of variance and stratification led to considerable inequality in a variety of educational outcomes for children from disadvantaged backgrounds, and b) empirical research which suggests that educational systems with high academic ability variance may, in reality, have poorer average performance. In relation to the former, Brunello and Checchi (2007) found that tracking is related to disadvantages for poorer children in both educational attainment and labor market outcomes, and that the effects are larger for earlier tracking. Jerrim et al. (2015) found that private schooling in Australia, the UK, and the US was associated with advantages in both education and labor market outcomes. Finally, Parker et al. (2016) found that ability stratification was associated with lower expectations of university attainment for poorer children net of academic achievement.

In relation to the second stream of evidence, Hanushek and Wößmann (2005) found that early tracking increased educational inequality and some evidence that it was associated with lower mean performance. Micklewright and Schnepf (2007) showed that the distance between the 95th and 5th percentile in achievement in a country and their median performance was negatively correlated. Likewise, Checchi et al. (2014) found no or negative relationships between various forms of variance and stratification and average achievement. In addition to data on educational performance is research on educational attainment. Thomas, Wang, and Fan (2001) found a negative relationship between a Gini index (a relative measure of inequality) of years of education and the average years of education within a country for rich countries. Pfeffer (2015) combined both research traditions to show that there is no trade-off between performance in international adult skills assessment and equity of opportunities. Overall, this suggests that there is little evidence of an excellence/equity trade-off in educational systems; at least within rich countries.

Almost all of the research to date, however, has focused on between (cross-sectional) rather than within (multi-cohort) country relationships. Likewise, it is also important to note that changes in stratification may or may not occur evenly across the achievement distribution, with changes in variance at the top or bottom half potentially being of most importance. Where changes in variance occur could potentially have different implications. Micklewright and Schnepf (2007) suggest, for example, inequality tends to be largest in the bottom of half of the achievement distribution. Thus, *increases* in polarity (movement from the median of the distribution to the tails) at the bottom end of the distribution may be most important. Indeed, Poland has had particular success at improving performance by targeting policy at such students (Breakspear, 2012). Alternatively, Ryan (2013), focusing only on Australia, suggests declines in the top half of the distribution account for that country's decline in math performance. This suggests that reduction in polarity at the top end of the distribution (i.e., the highest performers becoming more similar to the median performer) is of most concern.

Current Research

The current research makes use of over a decade of the Programme for International Student Assessment (PISA) data collection to explore the association between within country inequality and average achievement. We also use the multiple rounds to consider how within country changes in inequality and achievement are related over time. As such we advance the following hypotheses:

H1: Trends in inequality from 2000 to 2012 will be zero or negatively related to trends in performance over the same period.

H2: Changes in inequality from one PISA round to the next will be zero or negatively related to changes in performance. Both H1 and H2 are predicated on the hypothesis that there is no trade-off between excellence and equality in academic achievement.

H3: When large changes in inequality occur, changes in the top or bottom of the achievement distribution will be differentially associated with changes of average achievement.

Measures of Inequality

We note that a number of different measures of inequality have been used in the literature. These include measures focused on how children of different levels of ability are sorted into schools such as the between-school achievement variance or intraclass correlation coefficients (Marks, 2006; Parker et al., 2016; Salchegger, 2015). These measures provide an index of the degree to which a country's education system segregates children of different levels of academic ability into different schools and thus incorporates both formal differentiation (e.g., tracking) and informal (e.g., social segregation). Other measures focus on the degree of variance in academic performance between children within the same country. These include absolute or relative (i.e., scale invariant) measures (see Handcock & Morris, 1999 for a review). We use a selection of all of these indexes including a) intraclass correlations as a measure of the amount of between-school ability stratification (ICC; see Parker et al., 2016); b) the distance between the 95th and the 5th percentile in achievement as a measure of absolute variance in achievement (see Micklewright & Schnepf, 2007); and c) a constructed Gini index of achievement and, where possible, relative polarity as relative indexes of variance (Handcock & Morris, 1999).

There are few criteria for what indicates large or small variation in these measures. And this is particularly the case in the context in which we use them, where we rely on change over time. In the absence of criteria then, we utilise extensive sensitivity analyses using multiple measure, across multiple academic domains, and the use of multiple statistical methods. Thus, our focus is on results that show consistency across these approaches.

Method

Participants

All analyses were done at the country level using participant level indicators of math, science, and reading achievement from all five PISA rounds. We focus upon OECD countries (based on membership as of 2000) with the exception of Mexico and Turkey[1]. PISA provides data on a representative sample of 15-year-olds. The data is collected in a two-stage procedure with schools selected proportional to size and a random sample of 15 year olds selected from within each school (OECD, 2004). A set of weights is provided such that the sample is representative of the target population. In total, participants came from 27 countries with a total sample of 1,026,173 for analysis related to reading achievement and 957,735 for analysis related to math and science achievement. The reason for the difference in participant numbers is that all participants received the reading test in PISA 2000 but a sub-sample received either the math or the science tests. In all other PISA rounds all participants received estimated performance scores for all domains.

Measures

Academic Performance. Children's academic achievement was measured via performance on a standardized test in math, reading, and science. The achievement tests used in PISA are designed specifically to enable cross-national comparisons in academic achievement. PISA differs from other international measures of academic

[1] The use of OECD countries excluding Mexico and Turkey is relatively common (e.g., Mickelwright & Schepf, 2007; Parker et al., 2016). The reason for this is a) considerable differences between Mexico and Turkey and the rest of the OECD in GDP and human development indexes and b) a large number of not at school youth in these countries at the age of interest leading to potential systematic bias in estimates.

performance as it focuses on functional ability rather than knowledge or mastery of a curriculum. Answers from the achievement tests were summarized by the survey organizers into a single score for each of the three domains using an item-response model, the intuition being that true skill in each subject is unobserved and must be estimated from the answers to the test (see OECD, 2004, for further details). Five plausible values were generated for each pupil, estimating their true proficiency in each subject. These scores were scaled by the survey organizers to have a mean of 500 points and standard deviation of 100 across OECD countries in the first PISA round. Country average performance, Gini, and ICC estimates were all estimated with each of the plausible values separately and then averaged to provide country specific point estimates.

Gini Index. The Gini index was calculated separately for each academic domain, country, and PISA round. As with all measures used in the present research the Gini was calculated using the population weight via the reldist package in R (Handcock & Aldrich, 2002). The index varies between zero (indicating a uniform distribution of achievement) to one (indicating only a single individual had a non-zero achievement score). We multiplied the Gini index by 100.

Intraclass Correlation (ICC). ICCs estimate the degree to which students within a school resemble each other—and differ on average from those in other schools—in terms of academic achievement. Thus, higher estimates of ICCs reflect the degree to which schools are homogenous in the academic achievement. ICCs were estimated using the variance components taken from an ANOVA of achievement predicted by school membership and weighted by the population weight. We also multiplied these by 100 so that they varied from 0 to 100 (see Marks, 2006).

P95 – P5. The distance between the 95th and 5th percentile of achievement was likewise calculated after applying the population weight.

There was evidence of considerable change in achievement and all inequality indexes (see supplementary material).

Statistical Analysis

Modelling Approach. Hypotheses H1 and H2 relied on exploring the relationship between estimates derived for each country. We focus here on estimates derived using a series of multilevel models with PISA round estimates nested within country. As such all analyses are done at the country level, and no individual level data is modelled in the analysis reported in the results. There are debates about how appropriate the use of multilevel level models are in the context of country comparisons. In particular, there are concerns that random effects models remain common despite the fact that a) countries are rarely sampled randomly from a population (or in our case include all, or almost all, countries in a relevant population; i.e., the OECD) and b) country-specific estimates can be biased (due to shrinkage) when there are few countries (e.g., Byran & Jenkins, 2015). As such we also test the robustness of the results using country fixed effects models. Models were fit with random effects for country with inequality and PISA cycle estimated as fixed effects. Detailed consideration of model development is provided in the technical appendix.

For trajectory models multilevel growth curve models were estimated (H1). In each case both the intercept (i.e., initial level at year 2000) and slope (i.e., slope of linear interpolated trajectories from 2000 to 2012) were estimated as random effects for country. Such models were run separately for academic achievement and inequality measures. Country specific slope estimates were drawn from the resulting parameter estimates. We also calculated the simple difference between PISA 2000 and 2012 achievement and inequality measures and looked at the relationship between these. Growth curve models treat PISA cycle as an ordinal variable and thus summarize the change across PISA cycles in relation to, for example, achievement as a linear trend. The benefit of this is that it provides a simple summary measure that reduces the influence of noise around this trajectory, thus reducing the impact of outlier cycles (e.g., where a country experiences a notable increase in one but only one PISA cycle before returning to baseline levels).

It is possible, however, that these results may be biased as they impose a linear trajectory from PISA 2000 to PISA 2012. We aimed to account for this using change score models (H2). In this case achievement at round k+1 was regressed on achievement at round k with the regression estimate fixed to 1 (i.e., a simple difference score) and the change score of inequality from round k to k+1. The result of this specification was that change in achievement was predicted by change in inequality over the same lag. Random effects for country were included.

Variance Location. Hypothesis H3 focused on where changes in inequality occurred in the achievement distribution. Using the reldist package (Handcock & Morris, 1999) we isolated changes in the achievement distribution from 2000 to 2012 in relation to shape (e.g., changes in skewness) versus location (e.g., movements of the population as a whole up or down the achievement distribution). We used two approaches to this. First, we explored the relationship between relative polarity (RP, i.e., degree of movement from the median to the tails of the distribution from one PISA cycle to the next) and changes in achievement for all countries. Second, we selected several countries that displayed considerable change in achievement from 2000 to 2012 for a more detailed analysis. We use both RP measures as well as plots of changes in the achievement distribution, decomposed into location and shape changes. All RP indexes vary from -1 to 1, with negative values indicating decreased polarity or a general movement of values toward the median. The median relative polarity index (RPM) provides an overall estimate. This can be decomposed to explore the upper (URP) and lower (LRP) portions of the distribution.

Results

H1: Trajectory

We first looked at whether linear trends in achievement from 2000 to 2012 were related to linear trends in inequality. For this we extracted country level trends from a) a series of random intercepts and slopes models; b) a series of country fixed effect models; and c) the simple difference between achievement and inequality measures from PISA 2000 to PISA 2012 (hereafter simple). As can be seen from Table 1, the relationship between the trend in achievement and the trend in inequality was negative in all cases (including both Pearson and Spearman correlations). In support of H1, countries that increased in achievement from 2000 to 2012 tended to see a decline in inequality measures. Relationships were strongest for Gini and ICC indexes, with correlations routinely around -.50 and often above -.70. The relationships were more moderate for P95 – P5, and typically only significant for science. The correlations were similar for all achievement domains, with Figure 1 derived from the multilevel models, showing the relationship between the linear trajectory of math achievement and inequality. The technical appendix provides figures for reading and science.

H2: Change Scores

The above analysis focused on linear change in achievement and inequality from 2000 to 2012. It is possible, however, that these results do not give an accurate reflection of the relationship between changes in achievement and inequality. To account for this we looked at the relationship of changes in achievement regressed on changes in inequality from one PISA wave to the next (Table 2). For all academic domains, a change in the Gini index from one PISA round to the next was associated with a significant counteracting change in achievement levels. On their original metrics, a one-point increase in Gini (inequality) was associated with a 6 (for science) to 10 (for math) point decline in achievement. Put another way, a one-point increase in the Gini coefficient measure if inequality is associated with a 0.06 (science) and 0.10 (math) effect size decline in average performance. Significant relationships were likewise found for reading and science for ICCs and for reading for P95 – P5. Effect sizes were moderate for the Gini index and ICCs and small for P95 – P5.

H3: Where Does Inequality Change?

A focus on change scores also allowed us to consider changes in relative polarity from one PISA wave to the next. In all cases the estimates were negative suggesting that there was not a trade-off between excellence and equity in either the higher performing or lower performing students (see Table 2). Supporting H3, the effects for RPM and RPL were only significant in one case. Overall the relationships were strongest for the upper half of the achievement distribution and significant or marginally significant for all domains. This indicates that declines in achievement maybe more strongly weighted toward increases in inequality in the upper portion of the achievement distribution. Put simply, declining PISA scores tended to be associated with average performing students falling further behind the highest performing students; such that the right tail of the distribution became increasingly elongated (i.e., the highest performing students tend to be protected against declines in achievement). The difference between RPL and RPU were, however, relatively small – though nevertheless sufficient to suggest a more in-depth consideration would be beneficial.

We finally considered where in the achievement distribution changes in inequality tended to occur for countries that experienced notable changes. Given space constraints we focus here on Germany, Poland, Sweden, and Iceland as those countries in which the largest changes in achievement and inequality occurred. Germany and Poland were the only two countries to improve by over 20 achievement points and decreased in Gini by over one-point for each domain from 2000 to 2012. Sweden declined by almost 30 points in each domain and increased in Gini by well over one point in both reading and science (and over half a point in math). Likewise, Iceland increased in Gini by over one-point in each domain and declined in achievement by over 20 points in math and reading (and over 17 points in science).

The results indicate significant changes in polarity for each focussed country in at least two of the three achievement domains (see Table 3). Germany and Poland declined in polarity (see plots in supplementary material). Germany predominantly declined in the upper portion of the distribution with Poland displaying most change in the lower portion. However, for reading in Germany and reading and science in Poland significant declines in polarity occurred in both LRP and URP. This shape change resulted in fewer individuals in the lower and upper deciles than would have been the case if changes in achievement from 2000 to 2012 were due to location changes alone. Sweden and Iceland both significantly increased in RP. In both cases changes were predominantly located in the upper portion of the distribution. What this means is that, as Sweden and Iceland declined in average achievement, the most talented students were partially protected. Thus there were frequently 20 to 30% more students in the top decile than would be expected if achievement declines were consistent across the whole distribution. Indeed, for science achievement in Iceland there were approximately equal numbers of students in the top decile of the reference distribution at both PISA 2000 and 2012, when there should have been only 60% as many individuals in 2012 if there was no change in RP (see Figure 2).

Discussion

Consistent with growing evidence, our results suggest that inequality, indexed by stratification or variance in achievement, is negatively associated with average achievement at the country level. Importantly, effect sizes were routinely of a similar size for both relative Gini (variance) and ICC (stratification) indexes of inequality; though relationships were smaller, but still negative and often significant, for absolute measures of variance (see below). We extended previous research by focusing on within country changes in inequality and its association with within country changes in average achievement. Not only were within country results consistent with previous research in showing a lack of evidence for a trade off between excellence and equity, the current results suggested that inequality maybe associated with declines in performance.

Of further interest, when considered from a within-country perspective, traditional dividing lines between educational systems evaporated. In particular, while Nordic countries have often been shown to be among the most equal in between country studies (e.g., Parker et al., 2016, 2017), when considering within country estimates Iceland and Sweden had some of the most evident declines in achievement and increases in inequality of all countries considered. Alternatively, while Germanic countries have been shown to be some of the most unequal due to early and extensive tracking, Germany has shown considerable improvement in academic achievement and this has been associated with notable decreases in inequality. Taken together, while between-country differences continue to follow traditional demarcations in inequality – Nordic < Anglophone < Germanic – (see Dupriez & Dumay, 2006), within-country analysis shows a shifting landscape where these monikers hold less relevance. This could be taken to suggest that overall the inter-country landscape is becoming more equal. There were notable increases in ICC PISA 2000 to PISA 2012 and (see supplementary material). Thus, given these average increases in ICCs, it may be that the trend, for OECD countries at least, is toward greater inequality.

It may be that changes unrelated to direct educational policy are driving these results. As such, we ran further sensitivity analysis on the country fixed effects presented in Table 1. In this case we calculated the partial correlation coefficients between academic excellence and equity controlling for trajectories across the same period (2000-2012) in Gross Domestic Product (GDP in US dollars), average disposable income, and in the percentage of GDP spent on social welfare. As Table 4 shows the results were extremely similar to those reported in Table 1.

Why Is Excellence Not Positively Related to Higher Variance

A major question that emerges from the current research is why there is so little evidence for excellence/equity trade-off. To some degree this is answered by proponents of the trade-off argument themselves. Namely that decisions relating to the amount and type of education that a child should invest in is a decision not made by the child themselves but rather by parents or guardians. Such parents may not make decisions that lead to the best possible school placement (Friedman, 2002). Widespread and systematic inefficiencies in child assignment could account for the results noted here (see Pfeffer, 2015). Indeed, PISA data suggests misplacement occurs across the socioeconomic ladder (Parker et al., 2017). For example Maaz et al. (2008) note that in the Germanic system parents from well-off families often insure that their children are located in university track systems even when teacher recommendations are for lower track placements. Parker et al. (2017) note that children of richer parents pay for poor placement with decreased academic self-concept. Conversely, children of poorer parents may gain in self-concept by inaccurate school placement but pay in terms of more difficult pathways to university.

This would suggest the problem is not with the idea that a school system should tailor offerings to different levels of the achievement distribution but rather with its application in context. However, there may also be inherent problems that suggest issues may continue to occur even with perfect placement. Indeed educational psychology evidence points to a natural bias in the way young people form expectations. Marsh (2006) argues that children in more selective schools have lower academic self-concepts than they would have had they gone to more comprehensive schools; a so-called Big-Fish-Little-Pond effect. An important extension of this is that lower self-concept leads to lower performance in a reciprocal spiral (a reciprocal effects model [REM]; Marsh, 2006). This effect is larger in countries with more tracking or higher ICCs (Salchegger, 2015). It is possible that this bias in self-perceptions may account for some of the reason why more stratified systems do worse than expected if the excellence trade-off was apparent.

Alternatively, peer composition and its potential negative effect on the self-concept, and thus, their academic achievement (i.e., REM) of high performing students provide one mechanism to explain the current results. In addition, non-linear peer effects in learning quality likely provide equally compelling explanation of these results for the low end of the achievement distribution. Put simply, high performing students tend to lose very little from association with poorer performing students but that poorer performing students gain considerable benefits in terms of motivation and quality of peer interaction (Checchi, 2006; Hanushek, Kain, Markman, & Rivkin, 2003; Hanushek & Wößmann, 2005). From a policy perspective then, it may be that each country needs to determine whether students across the achievement distribution may actually benefits from more integrated classrooms; though always with an eye to the local policy context.

Changes in Inequality

The current research suggested that increases in stratification measures of inequality are associated with decreases in average achievement. We considered average change in variance for all countries, but also the form of this change. Declines in achievement were mostly associated with protection of high performing students and declines of average and low performing students. Taken as a whole, there was evidence of an effective hollowing out of the middle of the achievement distribution where there was increasingly polarization between the most talented students and the rest. More in-depth analysis of countries that changed considerably (i.e., 20 PISA points and 1-Gini point) provided a more nuanced perspective on this issue.

Ringarp and Rothland (2010) note that Sweden has moved from one of the most to one of the least centralized systems with increased school choice and privatization in the last few decades. Iceland has long had a decentralized school system with considerable school choice. However, decentralization was strengthened by policy in 2008 and the implication of decentralization likely increased after the global financial crisis where local communities responded to a reduction in educational funding in a diversity of ways (Ministry of Education, Science and Culture, 2014). Importantly this led to quite considerable regional differences in declines in PISA performance. In contrast, the ‘PISA shock’ of 2000 in Germany led to a national conversation on education, an increase in centralization and a focus on lower performers and immigrants (Breakspear, 2012). In Poland, there was a strong focus on the poorest performing students in response to PISA results (Breakspear, 2012). Our results suggest that for Germany increases mostly centred on the middle of achievement distribution such that more of the mass of the achievement distribution was located around the average. For Poland our results show the strong success of their focus on the bottom of the achievement distribution. Taking all results together, a hypothesis emerges that a countries' educational policy that mainly serves talented students will be associated with lower average performance, alternatively, a focus on the lower and middle portions of the achievement distribution leads to improvement. Overall there is a need for future research that focuses not just on changes in inequality overall but where changes occur and what implications this has for how a given country should determine its educational policy given its own unique context.

Measures of Inequality

It is worth noting that there were some modest differences in the results depending on the measure of inequality used. Before discussing such differences, however, it should be emphasised that there was a broad consistency. First, the direction of the relationship between inequality and performance was always negative regardless of the measure used or the model used to test the relationship. Second, each measure of inequality was significantly negative for at least one achievement domain in each model. Nevertheless there were differences. Primary among them was that the relative measures of variance (Gini) and stratification (ICC) were similar in size and routinely larger than the absolute measure (95th – 5th percentile). This may be due to the relative measures having proportional scale invariance while the absolute measures do not (Handcock & Morris, 1999). Given this property it maybe that the relative measures are more clearly comparable across time and context than the absolute measures.

Education Policy Consideration and Limitations Given the Current Evidence

Our research findings are consistent with a broader set of research (e.g., Checchi et al., 2014; Hanushek & Woesmann, 2005; Micklewright & Schnepf, 2007; Van de Werfhorst & Mijs, 2010) that has questioned the value of educational policies, at a state or nation level, that promote school differentiation and thus, there is a continued need to consider aspects of government policy related to decentralization, private or privatized schooling, and tracking. All these policies promote stratification by ability and as such do not appear to lead to higher average academic ability. Indeed, as noted above countries such as Sweden and Iceland have increased decentralization and school choice and have seen notable declines in performance, while Germany has moved toward increased centralization and seen an increase in performance. There are several consideration, however, that should be taken into account when interpreting what our results suggest for policy in a given country.

First, average PISA achievement is only one measure of an education systems performance, and it should be noted the achievement tests on which they are based are low stakes. Speaking against this is modelling which implies improvements in PISA scores are linked with real world outcomes such as economic growth (see Hanushek & Woessmann, 2010). Nevertheless, a wide range of outcomes should be considered along with the trade-offs between outcomes. For example, tracking maybe associated with poorer average achievement, however, retention though the full program of study is high suggesting that the cost of tracking in terms of average achievement may yet have benefits in terms of student completion (Checchi et al., 2014).

Likewise it should be noted that policies and social change at other levels of society may require an increase in decentralization and school choice, or at least make such policies more appealing. As Friedman (2002) has noted school choice maybe one of the only, or at least one of the most effective, means of reducing educational inequality in the face of increasing residential segregation by income by providing children in very poor regions access to high quality schooling. Indeed, countries like the US have seen exceptional increase in such segregation over the period of study in this research (Owens, Reardon, & Jencks, 2016) and thus there is good opportunity to test Friedman’s hypothesis. It should be noted, however, that initial empirical evidence suggests that school choice in the context of residential segregation may actually exacerbate inequality for disadvantaged children (Saporito, 2003).

It should also be noted that while the multi-cohort evidence presented here is a large step forward over previous cross-sectional evidence, the results should not be taken as causal. In particular, the causal direction is unclear. For example, our results show that a country can combine both excellence and equality with great success (see also Simola, 2005). However, it is not certain that equality leads to better performance or whether higher performance provides scope for countries to focus more closely on issues of equality. Likewise, the correlation between achievement performance and equality may be a proxy for other factors. In particular, social structure not school structure could drive these results; though previous research suggests this is unlikely (Dupriez & Dumay, 2006). More likely inequality in funding between schools or even regions within countries could account for these results (Owens, et al., 2016); likewise school-to-school or regional differences in school quality.

Age of First Selection and Other Challenges to Our Conclusions

A notable challenge to our interpretation of the results presented here is that they compare systems with different ages-of-first selection (Pfeffer, 2015). That is the age at which students are streamed into different tracks. Thus, for example, PISA considers students at age 15 and yet a number of OECD countries begin tracking students at age 16 (Bol et al., 2013; Pfeffer, 2015). This has particular implications for our interpretation of the results for Poland that, as part of the reform of the education system, lifted the age-of-first-selection from 15 to 16 years of age (Jakubowski et al., 2010). Thus, a criticism of these results is that systems that do not track before age 15 are merely delaying the inevitable. Indeed, Jakubowski et al. (2010) notes that consideration of Polish students after first selection at age 16 still dropped in achievement when compared to comparable students who continued in an academic track. There are several points to be made here. First, Pfeffer's results are similar in conclusion to ours despite focusing on the adult population. Namely, there appears to be little evidence of a trade-off between excellence (or quality in Pfeffer's terminology; see literature review) and equity even when both are measured in after schooling. Second, even if it is the case that inequality observed in differentiated systems would still eventually emerge in late tracking systems (Jakubowski et al., 2010), it is certainly not clear that this means that the achievement advantage that late tracking has over early tracking will dissipate completely. Again the consistency between our and Pfeffer's results would seem to indicate that this fear is unfounded.

Conclusion and Future Directions

This paper, in combination with a growing amount of cross-sectional empirical research, provides compelling evidence that a negative relationship exists between average achievement and inequality. This problematizes policies that promote decentralization, school choice, privatization, and segregation. Still research and theory needs to explain why this negative relationship exists and under what social conditions it holds. Furthermore there is clearly a need for research, which further evaluates how changes in variance at different points in the achievement distribution effect achievement. Put simply research needs to determine whether and when policies directed toward those in the bottom half of the distribution are most effective. Likewise there is a need to consider what forces are behind changes in variance over relatively short periods in some countries; noting that the current study covers only a single decade. In particular, in depth analysis of countries that have shown clear change are needed to unpack the various structures and polices that lead to increases or decreases in inequality.

Finally, there is a need for longitudinal versions of large-scale assessment such as PISA in order to determine long-term outcomes of equity and excellence. As a compromise, as PIAAC (the adult skills assessment version of PISA) develops, linking PISA and PIAAC in a synthetic panel design may prove advantages. Alternatively, assessments that incorporate a larger number of age groups and at different points in their schooling careers will be of importance to overcome difficulties associated with country differences in age of first selection. In particular, as Pfeffer (2015) notes, large-scale assessment which covers the final year of compulsory schooling would be beneficial. Nevertheless, utilising multiple cycles of PISA as we do here in focus attention on within country changes (where policy contexts tend to be less variant) provides a useful alternative.

References

Bol, T., & 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*, 285-308.

Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance*, OECD Education Working Papers, No. 71, OECD Publishing.

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

Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic policy, 22,* 782-861.

Bryan, M. L., & Jenkins, S. P. (2015). Multilevel modelling of country effects: a cautionary tale. *European Sociological Review*, *32*, 3-22.

Burger, K. (2016). Intergenerational transmission of education in Europe: Do more comprehensive education systems reduce social gradients in student achievement?. *Research in Social Stratification and Mobility*, *44*, 54-67.

Checchi, D. (2006). *The economics of education: Human capital, family background, and inequality*. Cambridge, UK: Cambridge University Press.

Checchi, D., van de Werfhorst, H., Braga, M., & Meschi, E. (2014). The Policy Response: Education. In Salverda, Nolan et al. (Eds). *Changing Inequalities and Societal Impacts in Rich Countries: Analytical and Comparative Perspectives* (pp.294-327) Oxford, UK: Oxford University Press.

Dupriez, V., & Dumay, X. (2006). Inequalities in school systems: effect of school structure or of society structure?. *Comparative education*, *42*, 243-260.

Friedman, M. (2002). *Capitalism and freedom*. Chicago, University of Chicago press.

Friedman, M. & Friedman, R. (1980). *Free to choose.* New York, N.Y.: Harcourt

Gans, J. & King, S. (2014). *Finishing the Job: Real world policy solutions in health, education, and transport.* Melbourne, Australia: Melbourne University Press.

Goldthorpe, J. H. (2007) *On Sociology*, 2nd ed. Stanford, CA: Stanford University Press.

Handcock, M., & Aldrich, E. M. (2002). Applying relative distribution methods in R. *Center for Statistics and Social Sciences. University of Washington*.

Handock, M.S. & Morris, M. (1999). *Relative distribution methods in the social sciences*. New York, NY: Springer.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement?. *Journal of applied econometrics, 18,* 527-544.

Hanushek, E.A. & Wößmann, L. (2005). Does educational tracking affect performance and inequity? Differences-in-differences evidence across countries. *The Economics Journal, 116,* 63-76.

Hanushek, E. A., & Woessmann, L. (2010). *The High Cost of Low Educational Performance: The Long-Run Economic Impact of Improving PISA Outcomes*. OECD Publishing. 2, rue Andre Pascal, F-75775 Paris Cedex 16, France.

Hoxby, C. M. (2003). School choice and school productivity. Could school choice be a tide that lifts all boats?. In C.M. Hoxby (Ed.) *The economics of school choice* (pp. 287-342). University of Chicago Press.

Huang, M-H. (2009). Classroom homogeneity and the distribution of student math performance: a country-level fixed-effects analysis. *Social Science Reseearch, 38*, 781–91.

Jakubowski, M., Patrinos, H. A., Porta, E. E., & Wisniewski, J. (2010). *The impact of the 1999 education reform in Poland*. Washington, DC: World Bank.

Jerrim, J., Chmielewski, A. K., & Parker, P. (2015). Socioeconomic inequality in access to high-status colleges: A cross-country comparison. *Research in Social Stratification and Mobility, 42,* 20-32.

Kerckhoff, A. C. (1995). Institutional arrangements and stratification processes in industrial societies. *Annual review of sociology*, *21*(1), 323-347.

Lucas, S. R. (2001). Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American journal of sociology, 106,* 1642-1690.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between‐school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives*, *2,* 99-106.

Marks, G. N. (2006). Are between-and within-school differences in student performance largely due to socio-economic background? Evidence from 30 countries. *Educational Research, 48*, 21-4.

Marks, G. N. (2013). *Education, social background and cognitive ability: The decline of the social*. Routledge.

Marsh, H. W. (2006). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology*. Vernon-Wall Lecture: British Psychological Society.

Mickelwright, J. & Schepf, S. (2007). Inequalities in industrialised countries. In S.P. Jenkins & J. Micklewright (Eds). *Inequality and poverty re-examined* (pp. 129-145). Oxford, UK: Oxford University Press.

Ministry of Education, Science and Culture (2014). Review of Policies to Improve the Effectiveness of Resource Use in Schools Country Background Report Iceland.

Nussbaum, M. C. (1998). *Cultivating humanity*. Harvard University Press.

Organisation for Economic Co-operation and Development [OECD] (2004). *PISA 2003 Technical Report.* Paris: Organisation for Economic Co-operation and Development.

Owens, A., Reardon, S. F., & Jencks, C. (2016). Income Segregation Between Schools and School Districts. *American Educational Research Journal,* 0002831216652722.

Parker, P. D., Jerrim, J., Schoon, I., & Marsh, H. W. (2016). A Multination Study of Socioeconomic Inequality in Expectations for Progression to Higher Education The Role of Between-School Tracking and Ability Stratification. *American Educational Research Journal*, Online First.

Parker, P.D., Marsh, H.W., Guo, J. Anders, J., Shure, N. & Dicke, T. (2017). An Information Distortion Model of Social Class Differences in Math Self-concept, Intrinsic Value and Utility Value. *Journal of Educational Psychology.* Online First.

Pfeffer, F. T. (2015). Equality and quality in education. A comparative study of 19 countries. *Social Science Research*, *51,* 350-368.

Raftery, A. E., & Hout, M. (1993). Maximally maintained inequality: Expansion, reform, and opportunity in Irish education, 1921-75. *Sociology of education*, *1,* 41-62.

Ringarp, J., & Rothland, M. (2010). Is the grass always greener? The effect of the PISA results on education debates in Sweden and Germany. *European Educational Research Journal*, *9,* 422-430.

Rowe, E. E., & Lubienski, C. (2017). Shopping for schools or shopping for peers: public schools and catchment area segregation. *Journal of Education Policy, 32,* 340-356.

Ryan, C. (2013). What is behind the decline in student achievement in Australia? *Economics of Education Review,37,* 226-339.

Salchegger, S. (2015). Selective School Systems and Academic Self-Concept: How Explicit and Implicit School-Level Tracking Relate to the Big-Fish–Little-Pond Effect Across Cultures. *Journal of Educational Psychology,* online first.

Saporito, S. (2003). Private choices, public consequences: Magnet school choice and segregation by race and poverty. *Social problems*, *50*(2), 181-203.

Simola, H. (2005). The Finnish miracle of PISA: Historical and sociological remarks on teaching and teacher education. *Comparative education*, *41*, 455-470.

Thomas, V., Wang, Y., & Fan, X. (2001). Measuring educational inequality: Gini coefficents of education. World Bank: Policy Research Paper 2525.

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

Walberg, H. J. (2000). Market theory of school choice. *Education Week*, *19*, 46-49.

Walzer, M. (1984). *Spheres of justice: A defense of pluralism and equality*. Basic Books.

Table 1

Correlation between Achievement Trajectory and Inequality Trajectory

	Random Effect (Pearson/spearman)	Country Fixed Effect (Pearson/spearman)	Simple (Pearson/spearman)
Math			
Gini	-.722***/-.727***	-.681***/-.756***	-.689***/-.725***
ICC	-.536**/-.519**	-.571**/-.563**	-.503**/-.457*
P95 – P5	-.347/-.374	-.313/-.369	-.323/-.439*
Reading			
Gini	-.613***/-.523**	-.567**/-.524**	-.542**/-.489***
ICC	-.495**/-.485**	-.558**/-.552**	-.670***/-.694***
P95 – P5	-.255/-.180	-.230/-.227	-.188/-.149
Science			
Gini	-.704***/-.727***	-.695***/-.662***	-.706***/-767***
ICC	-.440*/-.318	-.500**/-.414*	-.443*/-.281
P95 – P5	-.371*/-.464*	-.395*/-.392*	-.391*/-.518**

Notes. * *p* < .05; ** *p* < .01; *** *p* < .001. Random Effect = Correlation of slope with achievement slope from a multilevel growth curve model. Country Fixed Effect = Correlation of slope with achievement slope from a country fixed effect model. Simple = correlation of difference from PISA 2000 to 2012 in achievement and inequality measures.

Table 2
Lagged Results

	Est	SE	β	*p*
Math				
Gini	-10.556	1.908	-.369	***
ICC	-0.342	0.357	-.245	
P95 – P5	-.138	.075	-.104	^
RPM	-.454	.314	-.151	
RPL	-.178	.275	-.060	
RPU	-.562	.284	-.223	^
Reading				
Gini	-7.917	1.320	-.413	***
ICC	-0.531	0.178	-.458	**
P95 – P5	-.138	.059	-.165	*
RPM	-.454	.314	-.151	
RPL	-.614	.242	-.250	*
RPU	-.419	.252	-.172	^
Science				
Gini	-9.252	1.422	-.380	***
ICC	-0.499	0.183	-.380	**
P95 – P5	-.123	.072	-.127	^
RPM	-.706	.292	-.295	*
RPL	-.234	.278	-.100	
RPU	-.824	.242	-.397	***

Notes. ^p < .10 * *p* < .05; ** *p* < .01; *** *p* < .001. β = Estimates taken from a model in which achievement and stratification are standardized around the grand mean. Gini = Gini estimates of Achievement. ICC = Intra-class correlation of achievement, P95 – P5 = distance in achievement between the 95th and 5th percentile. RPM = Relative Polarity Median of achievement RPL = Relative Polarity Lower of achievement, RPU = Relative Polarity Upper of achievement.

Table 3

Relative Changes in Polarity

	Germany		Poland		Sweden		Iceland	
	Est	*p*	Est	*p*	Est	*p*	Est	*p*
Math								
RPM	-0.014		-0.059	***	0.017		0.072	***
RPL	-0.036		-0.117	***	-0.003		0.039	
RPU	0.009		-0.002		0.037		0.107	***
Read								
RPM	-0.093	***	-0.085	***	0.075	***	0.021	
RPL	-0.087	***	-0.115	***	0.063	**	0.017	
RPU	-0.099	***	-0.054	*	0.086	***	0.026	
Science								
RPM	-0.066	***	-0.092	***	0.026	*	0.086	***
RPL	-0.030		-0.095	***	-0.004		0.061	*
RPU	-0.101	***	-0.088	**	0.055	*	0.111	***

Notes. * *p* < .05; ** *p* < .01; *** *p* < .001. RPM = Relative Polarity Median of achievement, RPL = Relative Polarity Lower of achievement, RPU = Relative Polarity Upper of achievement.

Table 4

Country fixed effects controlling for country level covariates

	No controls (Pearson/spearman)	Social Welfare (Pearson/spearman)	GDP (Pearson/spearman)	Disposable Income (Pearson/spearman)[1]
Math				
Gini	-.681***/-.756***	-.680***/-.752***	-.694***/-.754***	-.710***/.775***
ICC	-.571**/-.563**	-.574**/-.557**	-.600***/-.562**	-.676***/-.656***
P95 – P5	-.313/-.369	-.313/-.381*	-.324/-.379	-.368/-.445*
Reading				
Gini	-.567**/-.524**	-.574**/-.536**	-.558**/-.551**	-.609***/-.519**
ICC	-.558**/-.552**	-.560**/-.549**	-.586**/-.540**	-.591**/-.544**
P95 – P5	-.210/-.264	-.210/-.271	-.226/-.281	-.327/-.311
Science				
Gini	-.695***/-.662***	-.696***/-.654***	-.753***/-.688***	-.785***/-.739***
ICC	-.500**/-.414*	-.503**/-.397*	-.536**/-.427*	-.553**/-.453**
P95 – P5	-.395*/-.392*	-.394*/-.379	-.496**/-.427*	-.558**/-.539**

Notes. * *p* < .05; ** *p* < .01; *** *p* < .001. All covariates were taken from the OECD (https://data.oecd.org/). [1] These results exclude Luxemburg for whom disposable income data was not available.

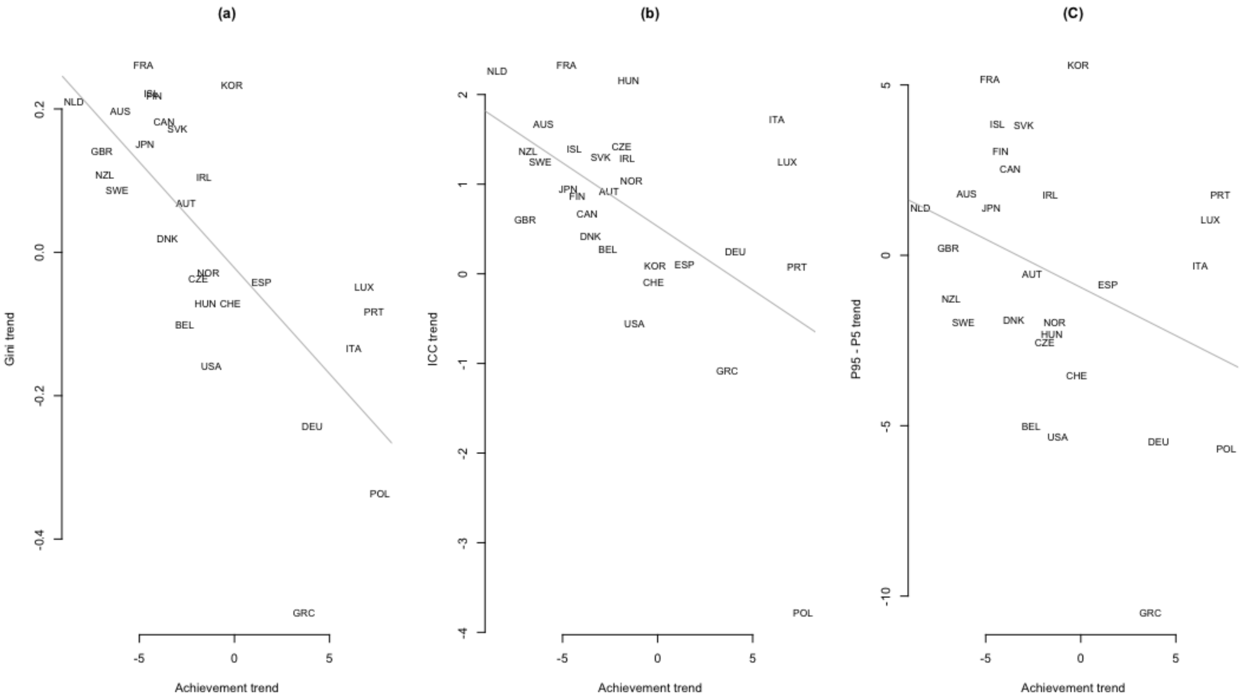

Figure 1. Math Trends.
Notes. Country given using the ISO 3-Letter code. Regression line represented in grey.

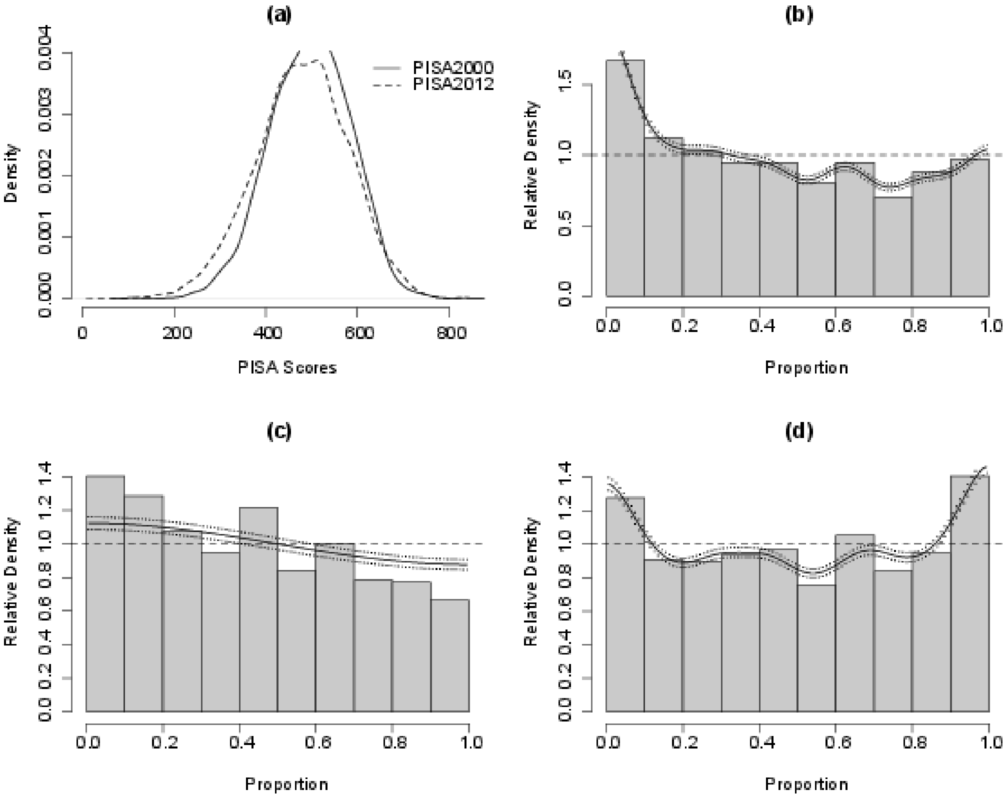

Figure 2. Distribution change for Science Achievement in Iceland.
Notes. Panel (a) represents the achievement distributions for PISA 2000 and 2012. Panel (b) represents changes in achievement from 2000 to 2012 using the 2000 distribution as a reference. Panel (c) indicates what the change in the achievement distribution from 2000 to 2012 if the change was due to location change alone. Panel (d) indicates the changes in variance at different points in the distribution. Bar plot indicate calculations of change within each decile of the achievement distribution. The solid line indicates estimates changes in the distribution using a Gaussian kernel density estimator. The dotted lines indicate 95% confidence intervals. For both the bar graph and estimated change line, a value above 1 indicates values in 2012 above that observed in 2000. Values below 1 indicate values in 2012 that are below those observed in 2000. Thus, when the confidence intervals do not cross the horizontal line at one on the y-axis indicates significant differences between 2000 and 2012.