Admissible statistics of educational achievement scores

Kristian Koerselman*†‡

March 30, 2010

Abstract

Labor economists regularly regress educational achievement scores on covariates to examine what affects achievement. I discuss the measurement and interpretation of achievement scores, and argue that, as the scores are typically measured on an ordinal scale, their analysis in terms of higher level statistics such as means is inappropriate, and that we should use quantile-based analysis instead. I investigate how large possible bias from mean-based methods is by comparing test score distributions to the distribution of monetary value of the same scores. In most cases, the bias will be quantitatively small, and conclusions qualitatively robust.

Keywords: *admissible statistics, educational achievement, item response theory, curriculum tracking.*

JEL: *C40,I20,I21,J24*

*Swedish Institute for Social Research SOFI, Stockholm, Sweden
†Department of Economics, Abo Akademi University, Turku, Finland
‡Contact information at http://economistatwork.com

1 Introduction

It is fairly common for labor economists to regress educational achievement scores on a wide range of covariates, – for example in order to evaluate educational policies – much in the same way as we regress wages and employment. While the latter two outcomes are relatively straightforward to interpret, educational achievement scores are not. What does a 13-point treatment effect on average test scores mean? And is a 20-point distance between two students at the bottom of the distribution the same thing as a 20-point distance at the top?

In some cases, test scores represent differences in the probability of succeding at a certain task. Equal score distances in different parts of the distribution reflect equal percentage point probability increases. These probabilities are however rather arbitrary: if we change the difficulty of the task, relative probabilities will change too. The distributional shape of the test scores is thus dependent on the questions we ask.

In other cases, score distances are the product of a more elaborate estimation framework called item response theory, or IRT. IRT ensures that score distances are not all too dependent on the level of questions asked. In this case, it is however the framework itself that imposes a distributional shape on the test score distribution.

As opposed to medians, means are sensitive to changes in distributional shape. Because of this, groupwise comparisons of mean scores are not robust to changes to the test process or score estimation procedure. This also holds for other mean-based methods, such as (mean-based) regression. In this paper, I argue that we therefore should use quantile-based methods instead, such as quantile regression.

We may believe that there is a true underlying concept of achievement score distances; it just happens that we are not able to identify those distances. One could assume that the underlying score distribution is normal, and indeed, by convention, educational achievement scores are made to have an approximate normal distribution. This is not entirely unreasonable. Many

physical and biological phenomena follow a normal distribution. There are however other 'naturally occuring' distributions as well.

As economists, we have an additional option. Educational achievement can be defined as its monetary value. When we look into this a bit further, we see that the monetary value of achievement appears to be lognormal rather than normal. The difference in distributional shape means that there can be cases where average test scores go up, but the average monetary value of those scores goes down, or vice versa. In the second part of this paper, I show under which circumstances such a qualitative reversal of conclusions can happen. If we consider the monetary value to be the true distribution, we can calculate the bias from using normalized data instead. I estimate the size of this bias using UK data, and conclude that a qualitative reversal should be uncommon, even in relatively extreme cases.

2 Admissible statistics

There are many statistics we could calculate from a particular data set, but not all make sense. All data are in essence mappings of empirical objects onto some scale or another. The choice of scale is to a certain degree arbitrary, and we would like our empirical conclusions to be invariant to the choice of reasonable alternative scales. For example, a comparison of mean heights of adult men in England and France should yield the same qualitative result whether measured in meters or in feet. In this case, the mean is indeed robust as the empirically taller nation will always have the higher mean height. By contrast, conclusions based on the mean of a nominal (or 'categorical') variable are not invariant to the choice of scale. Consider 'religion'. Using 1 for "Protestant", 2 for "Catholic" and 3 for "other" may or may not give a different ordering of mean(England) and mean(France) compared to using 1 for "Catholic" and 2 for "Protestant".

Going back to Stevens (1946), we can group scales into four levels: nominal, ordinal, interval and ratio (see Table 1). We call a certain statistic *admissible* for a level of scale when empirical conclusions derived from it are invariant

Scale	Mapping	Examples of variables	Examples of admissible statistics
Ratio (highest)	$x' = ax$	income, age	coefficient of variation
Interval	$x' = ax + b$	school grade (i.e. year), calendar date	mean, variance
Ordinal	$x' = f(x)$, $f()$ monotonically increasing	level of education, socioeconomic background	median, other quantiles
Nominal (lowest)	$x' = f(x)$, $f()$ gives a one-to-one relationship	gender, race, religion	mode

Table 1: Admissible statistics for four different measurement levels, adapted from Stevens (1946). Each measurement level inherits the admissible statistics from the levels below.

to the use different scales within the level. Statistics are always admissible on higher level scales than their own, and inadmissible on lower levels.

Of course, we are free to disregard this, and calculate inadmissible statistics anyway, as some do. The matter is perhaps not so much whether the statistic is admissible or not, but rather whether the statement we make on the basis of it is empirically *meaningful* (cf. Hand 2004, section 2.4.1). The statement "*The mean religion in France is 2.34.*"* does not have empirical meaning because there is no empirical counterpart to mean religion. It should be noted that even in this extreme case, the mean is not entirely void of empirical information (cf. Lord 1953, Zand Scholten and Borsboom 2009). If we find different means in two countries, we know that they differ in their religious compositions, even though comparing means is perhaps not the best method for identifying these differences. The statement 'the means are different' can be said to be meaningful even if 'the one mean is larger' is not.

There are cases where the underlying variable we are trying to measure is of the higher level even though the scale of the available data is not. In those cases, it may seem meaningful to formulate higher-level statements on the basis of the data, but our conclusions will not be robust to arbitrary changes

*Data: *Les Français et leurs croyances*, Le Monde, 2003

to the scale. If we want to make such statements nevertheless, we must add the higher level information by assumption, or from another source.

3 Measuring achievement

There are two main ways of estimating educational achievement. In Classical test theory or CTT, the score is based on the proportion of items answered correctly. This is the kind of scoring we perhaps remember from our own educational careers, often even extending into university exams.

CTT is based on a true score model

$$x = t + \varepsilon$$

where t is the true, underlying level of the student and x is the observed proportion of questions answered correctly. The error ε arises because the test procedure is noisy. Since we cannot ask the student infinitely many questions to find the true t, we use x as its estimate.

Test scores calculated using CTT are straightforward to interpret. The scores are estimates of the proportion of questions a student would be expected to answer correctly when given a similar test. Group averages of CTT scores also have a clear interpretation: the average score gives the proportion of questions the group as a whole would be expected to answer correctly.

The advantage of CTT is at the same time its disadvantage. CTT provides a score given a particular level of questions. The score distance between two students is determined by the level of questions considered. If the questions are very hard, almost no question will be answered correctly, student scores will be massed against the lower 0% bound, and consequently, the score distribution will have right skew (see Figure 1). Similarly, the score distribution will have left skew when the questions are very easy. In the first case, the score distances between low-scoring students become small, and between high-scoring students they become large. The opposite happens in the second case. (cf. Lord 1980, p. 50)

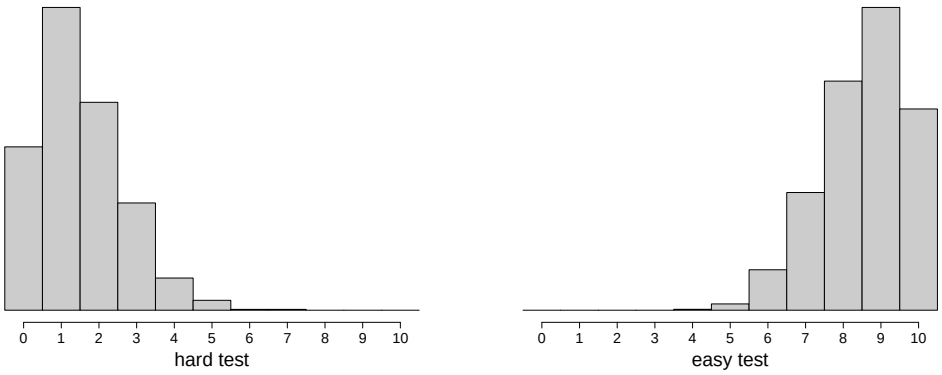

Figure 1: Hard CTT tests produce a score distribution with right skew while easy tests produce left skew.

An alternative to CTT is Item response theory, or IRT. IRT simultaneously estimates student and question properties by fitting a logistic *item response function*. For dichotomous questions (which are either answered correctly or not), the item response function is given by

$$\mathbb{P}(y_{ij} = 1) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}}$$

This function is illustrated in Figure 2. $\mathbb{P}(y_{ij} = 1)$ gives the probability of student i answering question j correctly (polytomous models are possible as well), θ_i is student achievement, b_j question difficulty, a_j question discrimination, and c_j is the limiting probability of answering the question correctly for extremely low levels of achievement. The upper probability limit is assumed to be one.

The inflexion point of the logistic curve lies at $b_j = \theta_i$, and we say that the student 'is of the same level as the question' at this point. The parameter a_j can be interpreted as the degree to which answering correctly on the question is related to the achievement dimension of the test, and c_j as the probability of guessing the correct answer.

There are model variations where one or more item parameters are fixed or otherwise restricted. When c is set to zero, and a to one, we obtain the common Rasch model. As is generally the case when $c = 0$, the inflexion point $b_j = \theta_i$ then lies at the level where the student is expected to answer the question correctly with probability 0.5.

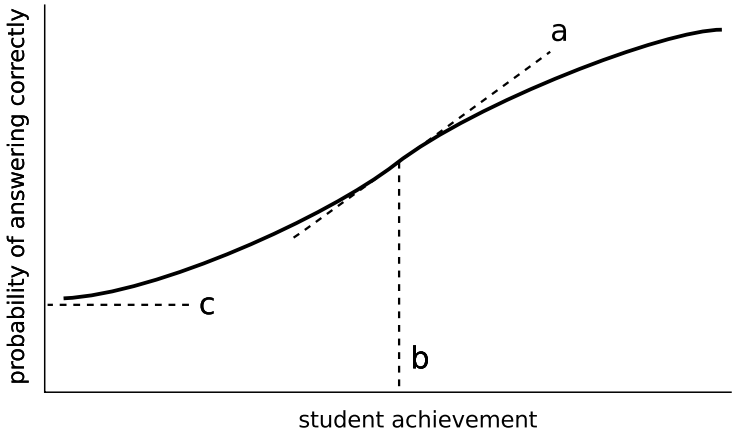

Figure 2: An item response function gives the probability of a student answering a certain question correctly as a function of his achievement. Question parameters a (discrimination), b (difficulty) and c (guessing) are illustrated in the figure.

Unlike CTT-scores, IRT student scores are not anchored to some absolute measure. We can for example add a constant to the vectors θ and b and arrive at the same model fit. In the same way, we could multiply θ and b with a constant and divide a by it. The model is therefore unidentified if we do not impose additional restrictions on the scores, for example by specifying that the sample mean score equals zero, and its standard deviation one.

In the IRT model, score distances arise from the difficulty with which students answer questions above and below their own level. If a student is answering questions above his own level with relative ease, $\theta_i - b_j$ must be relatively close to zero, just as when he does not do unusually well on questions below his level.

While IRT-estimated achievement distances are invariant to the choice of questions given to a particular student – the same estimated achievement

should arise from easy as from hard questions – they are not invariant to the way in which we estimate the model. We specifically estimate a logistic item response function function, and the model fits achievement to match this functional form. We could however just as well estimate a different item response function, and end op with another distributional form of achievement. The horizontal achievement axis can for example be transformed by $\theta^* = k_1 e^{k_2 \theta}$, where k_1 and k_2 are constants, so that both the item response functions and the achievement distributions are stretched out in one tail and compressed in the other (Lord 1980, p. 85).

4 How to deal with educational achievement scores?

Two questions arise when we consider how to deal with educational achievement scores. The first is a philosophical one: if we want to make meaningful statements on the educational achievement at an interval level, we must believe that there is an empirical counterpart to educational achievement distances. We must accept statements like 'George is as much better at math than John as Thomas is better than James': we must accept that going from a score of 20 to 40 is the same increase in achievement as going from 220 to 240. If we are not willing to do this, both comparisons of means and mean-based regressions are as meaningless as mean religion.

If we do accept that underlying educational achievement is of a higher level, we must ask ourselves if our estimation methods actually convey empirical information on that level. At first sight it could be argued that CTT and IRT scores so, and we may conclude that we can in fact use interval level statistics on the resulting scores. We should however realize that the information provided is not informative of an empirical concept we are interested in.

In the case of CTT, score distances are dependent on the difficulty of the test. If we compare mean scores between a treatment and a control group, we can draw an empirically meaningful conclusion on which group has a higher mean

score, but this result is not generalizable to CTT tests of different difficulty, and the result is therefore usually worthless from a policy perspective. Comparisons of means on the basis of IRT scores are robust to tests of different difficulties, but the shape of the score distribution is a result of the assumed functional form of item response curves. The shape of the distribution cannot be identified empirically, and the group with the higher mean under one set of assumptions may have a lower mean under different assumptions.

The most elegant way around these problems is to limit ourselves to ordinal information, and to compare quantiles (such as the median) instead of means. This solves both the philosophical interpretation problem and the practical measurement issues. If a certain educational policy (or another variable) has a positive effect on some parts of the distribution, but a negative effect on other parts, this should be reported. If it on the other hand has an uniform effect across all quantiles, this is worthy of reporting as well.

5 The real achievement distribution

Suppose that we believe in an underlying interval level interpretation of educational achievement, and that we persist in using mean based methods. How wrong can things go?

As I have argued, we cannot really measure interval-level information on the true achievement distributions, unless we want to limit our interpretation of achievement to an arbitrary feat, like the probability of answering questions correctly on a particular test. As economists however, we can interpret educational achievement at its monetary value. Monetary value is a variable of the ratio level, with a clear empirical counterpart to both the zero point and value distances.

The link between education and wages is of course not new. Economists regularly associate educational achievement with human capital (e.g. Becker 1964, 1993). Human capital is thought to improve the individual's productivity, akin to physical capital like tools and machines. Just as physical capital,

human capital needs investments for its creation, but unlike physical capital, human capital is embedded in its owner, and perishes with him when he dies. In this view, education is simply an institutionalized way to create human capital, and we can use the monetary value of education as a measure of its quantity.

How well then do educational achievement distributions fit productivity distributions? Not very well, as it turns out, at least not if we take wages as a proxy for marginal productivity. Educational achievement distributions are usually approximately normal. Wage distributions, on the other hand, are approximately lognormal, and when the width of the distribution is large enough, that can make quite a difference.

The left panel of Figure 3 shows the IRT-generated distribution of math achievement scores for the 2006 cross-section of 15-year old American children, collected by the Programme for International Student Assessment, or PISA (OECD 2006). Test scores are approximately normally distributed. The right panel shows US wages for full-time employed between 30 and 65 years old for 2004, as collected in the Luxembourg Income Study (LIS 2010). The wage distribution has clear right skew, and approximates the lognormal distribution. Indeed, the lognormal distribution generally fits country income distributions rather well, even though we may at times prefer the gamma or a three-parameter distribution. (Aitchison and Brown 1957, McDonald and Ransom 1979, Lopez and Serven 2006, Pinkovskiy and Sala-i-Martin 2009)

It is not hard to think of reasons why the wage distribution would have right skew. Just as normal distributions occur naturally as the sum of a large number of independent random draws (per the central limit theorem), lognormal distributions are generated by products of large numbers of independent (and positive) random draws. If percentual wage increases are drawn at random each day, the resulting wage distribution will be lognormal.

There are also other ways to explain right skew in wages, for example when high ability individuals educate themselves longer (Becker 1964, 1993, p. 100). Moreover, there is a difference between yearly wages and lifetime earnings: the yearly wage of the educated will have to be disproportionately higher to

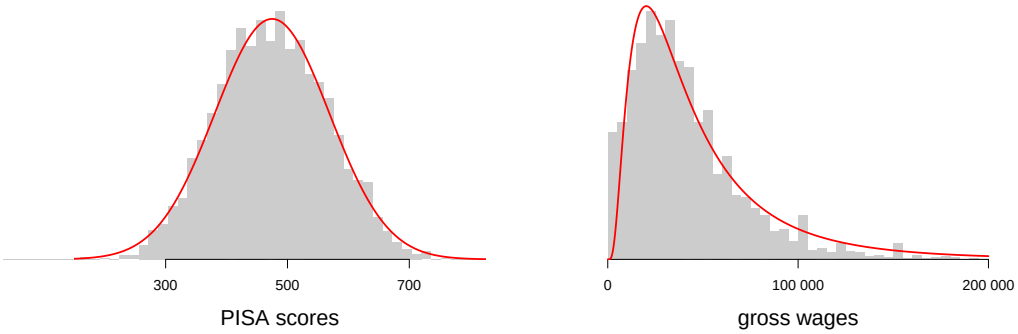

Figure 3: US PISA test scores follow an approximate normal distribution while US wages follow a heavily skewed approximate lognormal distribution. Shown: weighted US PISA educational achievement scores, 2006; US gross yearly wage income for full-time employed between 30 and 65 years old, 2004; fitted lognormal and normal distributions. Data: OECD 2006, LIS 2010.

compensate for the additional years of schooling; years which the could have spent working.

We do not actually want to compare raw distributions, but rather estimate the causal effect of an increase along the achievement distribution on wages for any point on the achievement distribution. Usually, we implicitly assume that the relationship between education and wages is loglinear, as for example in a Mincerian wage equation like

$$\log(wage) = \alpha + \beta \cdot educational\ attainment + \gamma \cdot experience + \delta \cdot experience^2.$$

In fact, it looks like the relationship between attainment and wages is even more convex than that (Belzil 2008).

The distribution of attainment can be more skewed than achievement if high-achieving students school themselves longer. The link between achievement and wages can however be estimated empirically as well. To do this, I take data from the longitudinal UK National Child Development Study (NCDS

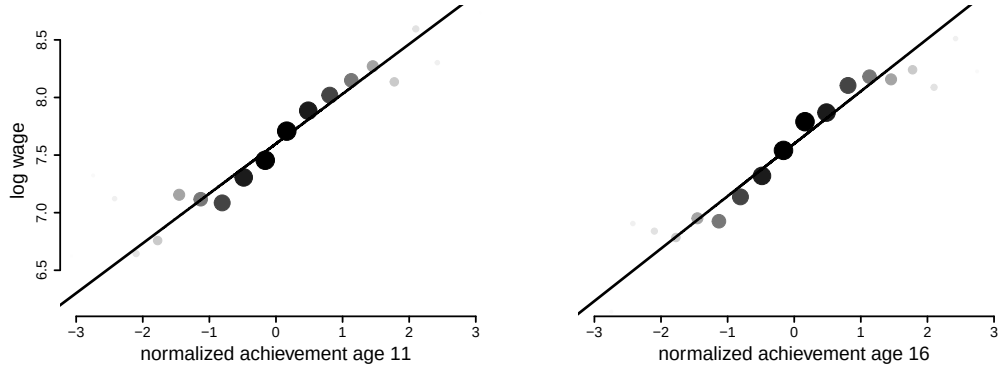

Figure 4: Average logged gross wages of 48-year old full-time employed males for different achievement levels (circles, circle area and color is proportionate to the number of observations) and the regression line through the unaveraged data. Data: NCDS 2010.

2010) and regress age 48 wages for full-time employed males on the first principal component of their age 11 and 16 achievement scores.

Figure 4 shows average logged gross wages for different achievement intervals at age 11 and 16 (circles), and the regression line through the unaveraged data. The loglinear model fits the data rather well. Of course, the variance of the logged wages as explained by achievement is smaller than the variance of raw wages. I add controls for socioeconomic background, and arrive at a conditional lognormal wage distribution with a logsd equal to 0.39 for the age 11 achievement distribution and 0.41 for the age 16 distribution. The estimated conditional wage distribution can be seen from Figure 5.

We can try to control for the most important omitted variable, ability, by including the first principal component of age 7 achievement. The estimates are then reduced to 0.32 and 0.33 respectively. It is not entirely clear whether we should want to do that though by including age 7 scores, we remove any effect of education before that age.

How should we handle the disparity in distributional shape between educational achievement and conditional wages? We could treat the two as separate

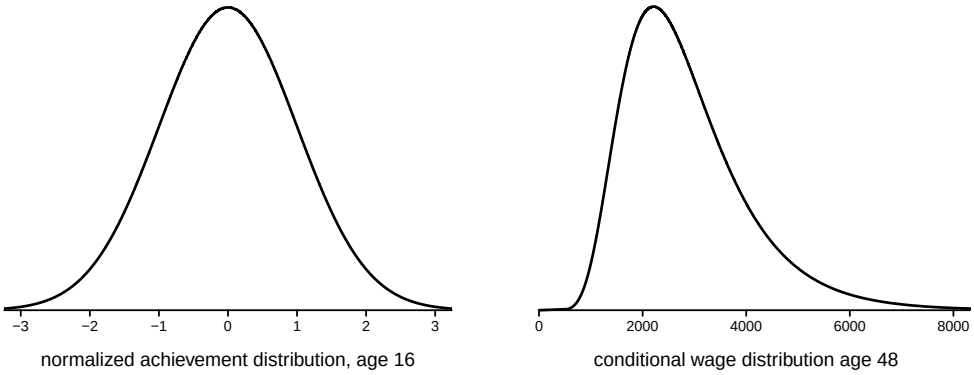

Figure 5: The estimated wage distribution conditional on differences in achievement levels, controlling for parental background. An one percentile in the achievement distribution (left) is associated with an one percentile increase in the wage distribution (right). Data: NCDS 2010.

concepts, and accept that the relationship between them is lognormal. This is not wrong per se. However, if we care about education because of future labor market outcomes, normal test scores are an inconvenient metric for those outcomes. For any single educational reform, average test scores may increase while average future wages decrease, or vice versa.

Under which circumstances can such a qualitative reversal of conclusions happen? Let us compare means between a treatment (subscript t) and a control group (subscript 0). I will call the difference between the two the treatment effect on the mean, or β_μ The true distribution is given by

$$y \sim \mathrm{LN}(\mu, \sigma^2),$$

while the normalized data are given by

$$y' = \ln(y) \sim \mathrm{N}(\mu, \sigma^2).$$

In order to catch only the change in the shape of the distribution, I will compare the difference of means in the normal distribution with the difference

of logged means (or logmeans) in the lognormal distribution. This means that the bias will be expressed in terms of the normalized test scores.

The estimate of the difference between the means β_μ is biased by:

$$\text{bias} = (\mathbb{E}[y'_t] - \mathbb{E}[y'_0]) - (\ln(\mathbb{E}[y_t]) - \ln(\mathbb{E}[y_0]))\,.$$

In terms of the moments of the treatment and control distributions, this equals

$$\text{bias} = (\mu_t - \mu_0) - \left(\mu_t + \frac{1}{2}\sigma_t^2 - \mu_0 - \frac{1}{2}\sigma_0^2\right) = \frac{1}{2}\left(\sigma_0^2 - \sigma_t^2\right).$$

In other words, the amount of bias generated by assuming a normal distribution where the lognormal distribution is appropriate depends on the difference in variance between treatment and control groups. A relatively smaller variance in the control group will lead to a negative bias of the treatment effect, and vice versa. I have illustrated this in Figure 6.

The dependence of qualitative robustness on the variance of the distribution can be generalised. Davison and Sharma (1988) show that mean differences between two normal distributions of equal variance are indicative of mean differences in any monotonic transformation of that distribution. It should also be noted that treatment does not have to be the cause of the observed heterogeneity in variance between treatment and control groups. Even if the groups may differ in their (conditional) variances for other reasons, the bias remains the same.

The next step is to calibrate this equation by plugging in an empirical σ_0 and σ_t. Let us assume that the distribution of our control group equals the reference distribution, and that the treatment distribution is proportionately wider.

$$\begin{aligned} \sigma_0 &= \sigma_{ref} \\ \sigma_t &= (1 + \beta_\sigma)\sigma_{ref} \end{aligned}$$

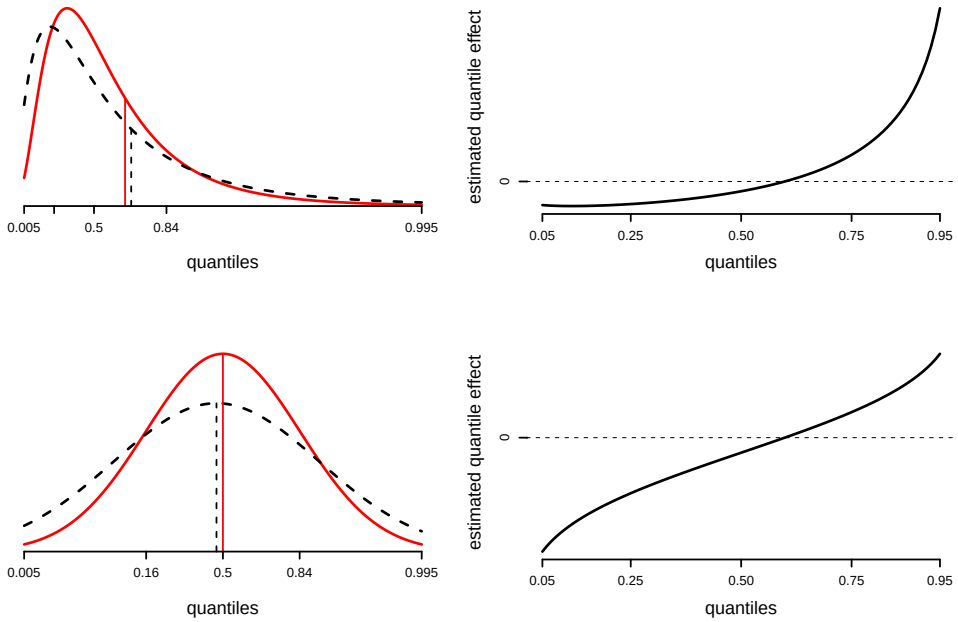

Figure 6: If the true distribution is lognormal (top left panel), normalizing data (bottom left panel) may lead to qualitatively wrong conclusions when comparing distribution means if the group variance differs. In this case, the treatment distribution (dashed lines) has a higher mean in the original data, but appears to have a lower mean after normalization. Median-based methods (right panels) are however quantitatively robust, with a negative effect on all quantiles below about 0.6, and a positive effect on all quantiles above.

Note that β_μ, β_σ and the size of the bias are now no longer expressed in absolute units, but in standard deviations of the original data. By substituting in the above equations, we arrive at a new expression for the bias in terms of the shape of the reference distribution.

$$\text{bias} = -\sigma_{ref}^2 \left(\beta_\sigma + \frac{1}{2}\beta_\sigma^2 \right)$$

Paper	β_μ	β_σ	σ_{ref}	bias	corrected β_μ
Hanushek and Woessmann (2006)	-0.179	0.101	0.41	-0.018	-0.161
Pekkarinen et al. (2009)	-0.007	0.009	0.41	-0.001	-0.006
Duflo et al. (2008)	0.175	0.042	0.41	-0.007	0.182

Table 2: Estimated treatment effects of curriculum from a number of selected papers, corrected for distributional form in the last column.

How large is the bias in practice? In many cases, variances are more or less constant over treatment levels, and the bias will be close to zero. There are however notable exceptions.

One such exception is curriculum tracking, the separation of students into different schools or classes based on ability. Such stratification almost certainly leads to larger differences between students (Koerselman (forthcoming), Pfeffer(forthcoming)). I have selected three empirical papers from the literature on the subject for further analysis.

Hanushek and Woessmann (2006) compare tracking policies between countries cross-sectionally on the basis of PISA/PIRLS and TIMSS data. Pekkarinen et al. (2009) investigate the effect of the 1970s Finnish comprehensive school reform using panel data, while Duflo et al. (2008) use a randomized trial in Kenya to look at the effects of tracking. These are three quite different settings, and their respective results are not necessarily generalizable across regions and times. It is therefore perhaps not surprising that the three papers find significant effects on the mean of different signs. Tracking is associated with larger differences between students in all three papers.

The first (numerical) column in Table 2 shows standardized estimated treatment effects on the mean from these papers. The second column contains the effects on the distributions' standard deviations. In the case of Pekkarinen et al. (2009) and Duflo et al. (2008), the effects on the standard deviations are not explicitly listed in the papers, but I have instead calculated them from other available statistics.

As a rough back of the envelope estimate of the robustness of these tracking estimates, I apply the logsd of the conditional UK wage distribution from

Figure 5 to the test score distributions. The size of the resulting bias as well as corrected estimates can be found in the last two columns of the table.

The size of the bias is quantitatively small; under 0.02 of a standard deviation in test scores for all three papers. This is not enough to change the papers' respective qualitative conclusions, which is encouraging. I have also made an effort to match wage distributions of the respective papers' geographical areas using data from the WIDER World Income Inequality Database (2010), the Penn World Table (Heston et al. 2009), and the Luxembourg Income Study (2010). The results are quite similar, and are not reported here.

6 Conclusions

If we want to make mean-based comparisons of educational achievement, we must make two assumptions. On a philosophical level, we must judge that the concept of score distances has an empirical foundation: score distances must be comparible in different parts of the distribution. Even if we believe that this is the case, the practical problem that measured score distances must accurately reflect distances in the underlying achievement remains. Whatever our answer to the philosophical question, measured distances are not robust to arbitrary changes in the choice of test in any case, meaning that comparisons of means are not robust either. An elegant and straightforward solution to both problems is to use quantile-based analysis instead, such as median regression.

Test scores commonly follow an approximate normal distribution. This shape is arbitrary, but it may nevertheless approximate the true underlying distribution of educational achievement. As economists, we can interpret educational achievement at its monetary value, making score distances comparable in terms of money. The monetary value of educational achievement is however approximately lognormally distributed. This is not usually a large problem: the size of the bias resulting from the use of normalized data is quantitatively small, and unlikely to lead to qualitatively different conclusions, even for known inequality-increasing policies like curriculum tracking. A note of

caution is in place here: the size of the bias is quadratic in both the logsd of the reference distribution and the difference in widths between the treatment and control distributions, and it increases rapidly for parameter values higher than those reported here.

Of course, even if we feel that we can meaningfully use means of educational achievement scores to evaluate educational policies, we still have preferences on the entire shape of the distribution. Applying a social welfare function to achievement distributions will tend to put a larger weight on the outcomes in the lower tail, and we may prefer outcomes with lower inequality even if this it associated with a lower mean as well. This argument should however not be hidden from the reader by changing the assumed educational distribution in a way that makes a change in the the apparent mean reflect changes in the assumed social welfare by normalizing data.

Acknowledgments

I thank Markus Jäntti, Denny Borsboom, René Geerling and Annemarie Zand Scholten for their kind help and advice. I gratefully acknowledge financial support from *Stiftelsens för Åbo Akademi forskningsinstitut*, *Bröderna Lars och Ernst Krogius forskningsfond*, *Åbo Akademis jubileumsfond*, and from the *Academy of Finland*.

References

J. Aitchison and J.A.C. Brown. *The Lognormal Distribution*. Cambridge University Press, 1957.

Gary Becker. *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press, 1964, 1993.

C. Belzil. Testing the Specification of the Mincer Wage Equation. *forthcoming in Annals of Economics and Statistics*, 2008.

M.L. Davison and A.R. Sharma. Parametric statistics and levels of measurement. *Psychological Bulletin*, 104(1):137–144, 1988.

Esther Duflo, Pascaline Dupas, and Michael Kremer. Peer effects and the impact of tracking: Evidence from a randomized evaluation in kenya. NBER Working Paper No. 14475, 2008.

David Hand. *Measurement theory and practice*. Oxford University Press, 2004.

Eric Hanushek and Ludger Woessmann. Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, 116:C63–C76, 2006.

Alan Heston, Robert Summers, and Bettina Aten. Penn World Table 6.3. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, 2009.

H. Lopez and L. Serven. A normal relationship? Poverty, growth, and inequality. *World Bank Policy Research Working Paper 3814*, 2006.

Frederic Lord. On the statistical treatment of football numbers. *American Psychologist*, 8:750–751, 1953.

Frederic Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum, 1980.

Luxembourg Income Study (LIS). Micro database, 2010; harmonization of original surveys conducted by the Luxembourg Income Study asbl. Luxembourg, periodic updating. 2010.

J.B. McDonald and M.R. Ransom. Functional forms, estimation techniques and the distribution of income. *Econometrica*, 47(6):1513–1525, 1979.

National Child Development Study (NCDS). National Child Development Study 1958–. 2010.

Organization for Economic Co-operation and Development OECD. Programme for International Student Assessment PISA. 2006.

Tuomas Pekkarinen, Roope Uusitalo, and Sari Kerr. School tracking and development of cognitive skills. VATT working paper 2, 2009.

M. Pinkovskiy and X. Sala-i Martin. Parametric Estimations of the World Distribution of Income, 2009.

Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

World Institute for Development Economics Research of the United Nations University UNU-WIDER. World Income Inequality Database WIID2b. 2010.

Annemarie Zand Scholten and Denny Borsboom. A reanalysis of Lord's statistical treatment of football numbers. *Journal of Mathematical Psychology*, 53:69–75, 2009.