40

Admissible statistics of educational test scores

Kristian Koerselman

Swedish Institute for Social Research SOFI, Stockholm, Sweden

Department of Economics, Abo Akademi University, Turku, Finland

Admissible statistics of educational test scores

Kristian Koerselman

Swedish Institute for Social Research SOFI, Stockholm, Sweden

Department of Economics, Abo Akademi University, Turku, Finland

Economists regularly regress educational achievement scores on covariates, for example to evaluate educational policy. I discuss the measurement and interpretation of achievement scores, and argue that, as the scores are typically measured on an ordinal scale, the use of mean-based methods such as OLS is inappropriate, and that we should use quantile-based analysis instead. Results based on regression are not robust to changes in test score estimation assumptions and methods. I investigate how large possible bias from mean-based methods is by comparing results using normal test score distributions and the lognormal wage distribution conditional on the same scores respectively. In most cases, the bias will be quantitatively small, and conclusions qualitatively robust.

Acknowledgments: I thank Markus Jäntti, Denny Borsboom, René Geerling, Annemarie Zand Scholten and Christer Gerdes 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*.

1 Introduction

It is common for economists to regress educational achievement scores on a wide range of covariates, much in the same way as we regress wages and employment. Common types of regression analysis, such as OLS, are effectively comparing means between distributions. Taking means of achievement scores poses two methodological problems.

First of all, we must believe that means of achievement scores are informative of the empirical world. This is not necessarily the case. Consider the following two statements.

(1) *The mean religion in France is 2.34.*

(2) *The mean salary in France is €25000.*

If we use the values 1 for “Protestant", 2 for “Catholic" and 3 for “other", we we can certainly compute the mean of these numbers and arrive at a figure of 2.34 for France. It is however clear that the first statement bears little relationship to the empirical world, while the second one is much less problematic. What is not clear is in which category achievement should fall.

Whether one thinks that there exists such a thing as mean achievement or not, we are still left with a problem of robustness. We cannot measure achievement at a higher level than the ordinal. The mapping of ranks to point scores involves implicit or explicit assumptions on the true distributional form of educational achievement. If we change those assumptions, the distributional form changes, and with it possibly our qualitative statements. If we compare mean test scores of two groups, the group that has the higher mean test score may have the lower mean under different assumptions.

While true achievement could have any distributional shape in theory, some shapes are perhaps more reasonable than others. Often, the underlying distribution of achievement is assumed to be normal, perhaps because many physical and biological phenomena follow a normal distribution. Normal distributions are usually the result of an additive process, in which each observed value in the distribution is the result of a repeated addition of small, independent random draws.

There are however other kinds of 'naturally' occurring distributions. If we believe achievement to be the result of a multiplicative process in which individuals learn at a randomly drawn rate every day, the resulting distribution will be lognormal. A multiplicative process implies that the amount of new learning is correlated with the amount previously learned: the children who have managed to achieve the most up until today, should be expected to learn the most tomorrow in absolute terms. If the process is additive, past and future learning is uncorrelated.

We do not necessarily have to derive the true shape from theory. Instead, we can link achievement to another, known distribution. Seen from a human capital perspective, educational achievement is a production input, and has a market value which can be estimated empirically. When we look at the shape of the wage distribution conditional on test scores, we find that it takes a lognormal shape.

As a kind of robustness check, we can compare regression results under the assumptions that achievement is either normal or follows the lognormal conditional wage distribution. I derive an expression for the difference in estimated treatment effect under the two conditions, and calibrate it with an estimate of the conditional wage distribution. Even if quantile-based methods are a more elegant way to handle educational achievement scores, mean-based methods turn out to be relatively robust in most cases.

2 Admissible statistics and meaningfulness

Psychometricians have a long tradition of linking appropriate statistical methods to different kinds of data. A key insight is that all data are in essence mappings of empirical phenomena onto some scale or another, and that the choice of scale is to a certain degree arbitrary.

We calculate statistics from our data in order to learn something about the real, empirical world. Statements on the data which bear no relationship to the empirical world, are therefore not *meaningful* (cf. Hand 2004, section 2.4.1). Statement (1) above does not have empirical meaning because there is no empirical counterpart to mean religion. While we can technically calculate the mean of an appropriately coded variable *religion*, doing so seems futile.

Apart from being meaningful, we e also want our statements to be robust to changes in the mapping from the empirical world onto the data scale. For example, we do not want our qualitative conclusions to change when we map height into meters instead of feet. A comparison of mean heights of adult men in England and France should yield the same qualitative result in either case. Comparing mean height is indeed robust as the empirically taller nation will always have the larger mean height. By contrast, conclusions based on the mean of a nominal (or 'categorical') variable are not robust 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 the English and French means compared to using 9 for "other" instead of 3.

Stevens (1946) suggests a relatively easy way to determine when we will run into robustness problems of the above kind. We group scales into four levels: nominal, ordinal, interval and ratio, as can be seen from Table 1. We call a certain statistic *admissible* for a level of scale when empirical conclusions derived from it are robust to the use different scales within the level. Statistics are always admissible on higher level scales than their own, and inadmissible on lower levels.

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

Scale	Mapping	Examples of variables	Examples of admissible statistics
Ratio	$x'=ax$	income, age	coefficient of variation
(highest)			
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

Meaningfulness and admissibility usually coincide, but there may be situations in which they do not (cf. Lord 1953, Zand Scholten and Borsboom 2009). We could for example compare mean

religion in England and France, and conclude that they are significantly different: that the English and French samples are likely not to have been drawn from the same population. The existence of a difference of the calculated means is dependent on the coding of the variable, and thus not robust, nor is the mean the best way to quantify this difference, but the conclusion that the religious composition of the two countries differ is meaningful nevertheless.

3 Dealing with achievement scores

Economists usually seem to assume that achievement can be measured directly, like physical measures of height or weight. Achievement must however be estimated, usually from the results of an achievement test. There are two methods of estimating achievement, but we cannot measure achievement at a higher level than the ordinal with either.

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 youth.

CTT is based on a true score model

$$x=t+\varepsilon$$

where t is the true, underlying achievement 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. We could thus conclude that CTT scores are of ratio level, and we would be right to do so, if there were just one possible relevant test.

The advantage of CTT is however 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)

While we can interpret CTT scores on a ratio level when speaking about a specific test, doing so precludes us from generalizing the result to the scores obtained by a different test. If we

want to make statements about generalized, underlying achievement as opposed to the ability to do a specific test, we must thus treat CTT scores on the ordinal level.

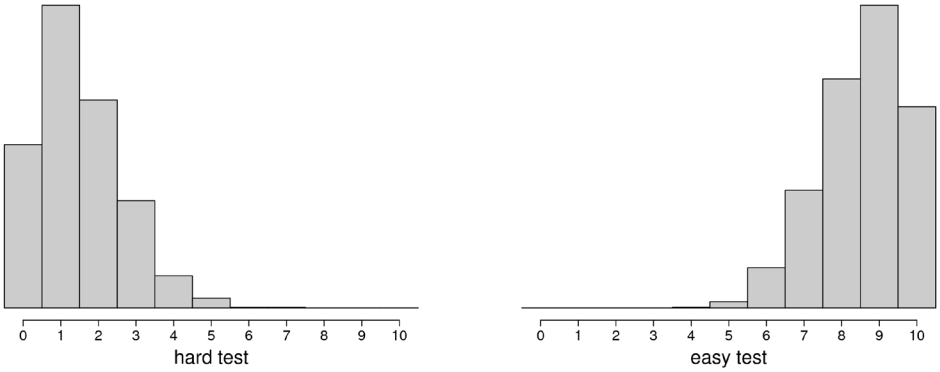

Figure 1: Hard CTT tests produce a score distribution with right skew while easy tests produce left skew.

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

$$P\left(y_{ij}=1\right)=c_j+\frac{1-c_j}{1+\exp\left(-a_j\left(\theta_i-b_j\right)\right)}$$

This function is illustrated in Figure 2. $P\left(y_{ij}=1\right)$ 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 student achievement and question difficulty are equal 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 then lies at the level where the student is expected to answer the question correctly with probability 0.5.

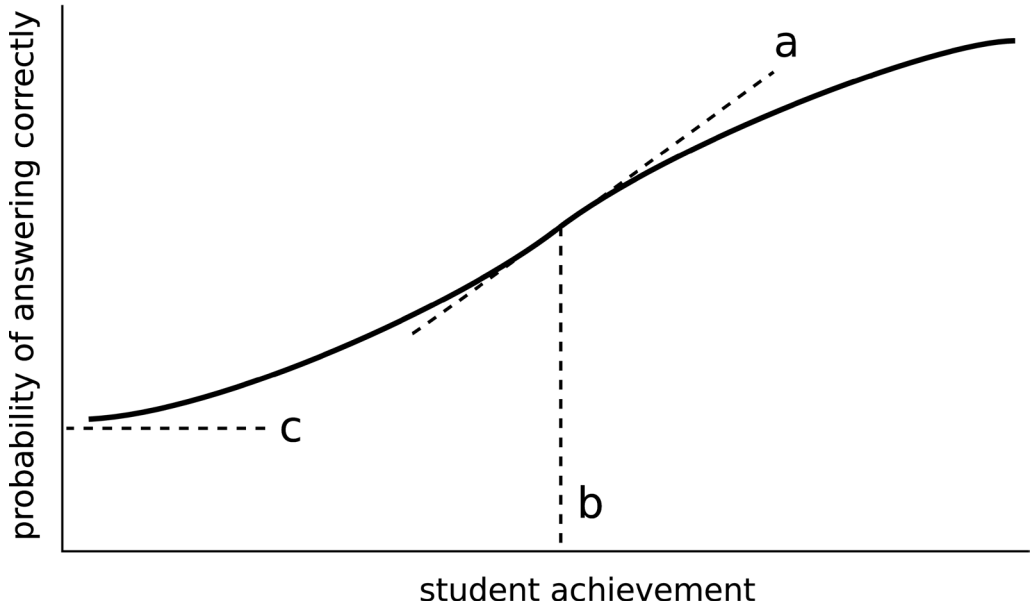

Figure 2: An item response function gives the probability of a student answering a certain question correctly as a function of his achievement. Question parameters *a* (discrimination), *b* (difficulty) and *c* (guessing) are illustrated in the figure.

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 of achievement with relative ease, 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 robust 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 robust to the way in which we estimate the model. We specifically estimate a logistic item response function function, and the model fits item parameters and 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 and difficulty axis can for example be transformed by $\theta' = k_1 \exp(k_2 \theta)$, where k_1 and k_2 are constants, so that both the item response functions and the achievement distribution are stretched out in one tail and compressed in the other (Lord 1980, p. 85).

Where CCT distances are a product of the particular test taken, IRT distances depend on the estimation assumptions given raw test scores. In both cases, we can reasonably change our

methods, and obtain a different test score distribution. In both cases, it is imprudent to interpret the scores on a higher level than the ordinal.

Given what we know about admissible statistics, meaningfulness, and the process by which test scores are estimated, how should we deal with educational achievement scores? Regression analysis is the comparison of means conditional on the relevant treatment status, and on other variables. The meaningfulness of regression is thus equal to the meaningfulness of a comparison of means.

The question of meaningfulness boils down to whether we believe that there exists an empirical phenomenon of underlying interval level achievement. If there exists such a thing, score distances must be comparable across the distribution. We must for example accept statements like

When it comes to math, Adam is as much better than Bert as Charlie is better than Dave.

Suppose that Adam can solve questions involving logarithms, Bert square roots, Charlie multiplication, and Dave addition. Is it meaningful to say that the difference between logarithms and square roots is just as large as the difference between multiplication and addition?

I will leave it to the individual empiricist to decide on the meaningfulness of a comparison of test score means. We should however be aware that comparing means and running regressions implies that we think that the above statement makes sense – that score distances are comparable throughout the distribution.

Even if we accept the meaningfulness of mean achievement, we are still left with the problem of measurement. As we have seen, estimated score distances are to a certain degree arbitrary. A comparison of means will have robustness problems in line with the theory of admissible statistics.

Both the problem of meaningfulness and the problem of admissibility can be solved by using statistics of the correct level. Instead of comparing means, we can compare medians, and instead of ordinary regression we can use median regression or the more general quantile regression. Doing so is robust as well as elegant in the sense that it fits to the level of information we can observe empirically.

4 The real achievement distribution

While I advocate the use of quantile-based methods, at the very least as a robustness check, it seems useful to to get some grips on just how large robustness problems are when using mean-based methods. Even if we cannot measure the distributional shape of educational achievement, we can try to make multiple reasonable assumptions on that shape, and look at how much results vary between them.

Two distributions stand out as natural candidates for educational achievement, the normal distribution and the lognormal. The normal is a common test score distribution, and test makers sometimes actively tweak tests to yield a normal distribution. It has theoretical appeal, as it emerges naturally from an addition of many independent draws from an arbitrary, finite distribution per the central limit theorem. The lognormal distribution however also has a relationship to the central limit theorem. If we multiply rather than add the (positive) draws, we will end up with a lognormal distribution.

If we want to justify the use of normal or lognormal distributions for educational achievement distributions through the central limit theorem, we must think of learning as a process in which students start from the same baseline, and learn small, random amounts each day, finally arriving at their test-day achievement level. A normal distribution implies that we think of learning as an additive process, where each new addition of knowledge is independent of previous draws. A lognormal distribution implies a multiplicative process, with independent draws of learning *rates*, but correlated learning *amounts*, such that higher achieving students are expected to acquire additional knowledge in the future than their peers.

There are other reasons which make the lognormal distribution appealing. Even if learning would be additive in principle, and innate ability would be normally distributed, the achievement distribution will have right skew if high ability individuals put more effort, time or other resources into learning (cf. Becker 1964, 1993, p. 100). In this light, if the eventual achievement distribution is to be normal, the distribution of innate ability must have left skew.

There is a third argument for a lognormal distribution of achievement. We can interpret educational achievement at its monetary value on the labor market. 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. 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 output.

What does the relationship between normally distributed test scores and wages look like? I take data from the longitudinal UK National Child Development Study (NCDS 2010) and regress age 48 wages for full-time employed males on the first principal component of their normalized age 11 and 16 achievement scores. Figure 3 shows average logged gross wages for different achievement intervals at age 11 and 16 (circles), and the regression line through the unaveraged data. There seems to be a linear relationship between test scores and the log of wages, which implies an exponential relationship between scores and wages. If we therefore map a normal score distribution into a conditional wage distribution, the latter should be lognormal.

Adding controls for socioeconomic background, I arrive at a conditional lognormal wage distribution with a logsd[1] equal to 0.39 for the age 11 achievement distribution and 0.41 for the age 16 distribution. The estimated conditional wage distribution for age 16 scores can be seen from Figure 4.

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 as we remove any effect of education before age 7 by including scores at that age. Also, we can keep in mind that by leaving out controls positively related to both achievement and wages, we will arrive at an overestimate of the causal effect of achievement scores on wages, meaning that the robustness check will be more conservative than it would otherwise be.

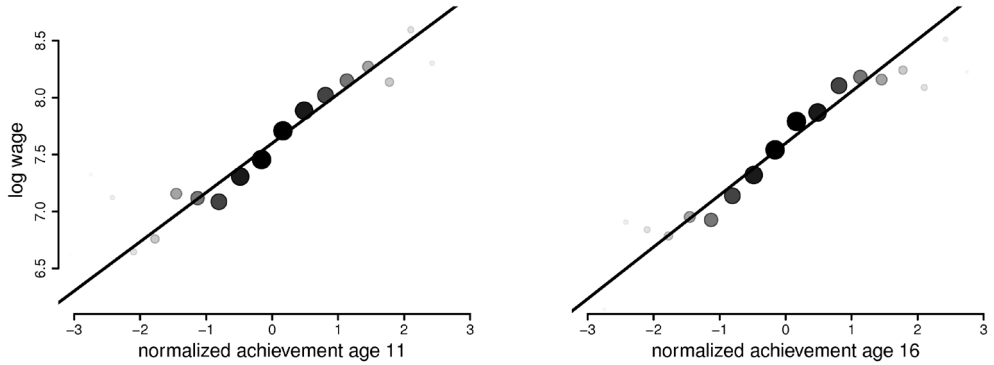

Figure 3: Average logged gross wages of 48-year old full-time employed males for different achievement levels (circles, circle area and color is proportionate to the number of observations) and the regression line through the unaveraged data. Data: NCDS 2010.

Having selected the default normal distribution and a fitted lognormal distribution as reasonable candidates for the true distribution of educational achievement, how much will regression results vary between frameworks where we assume either distribution?

[1]The logsd is the standard deviation of the logged values of the lognormal distribution. A lognormal distribution can be fully described by its logmean: the location parameter, and its logsd: its shape parameter.

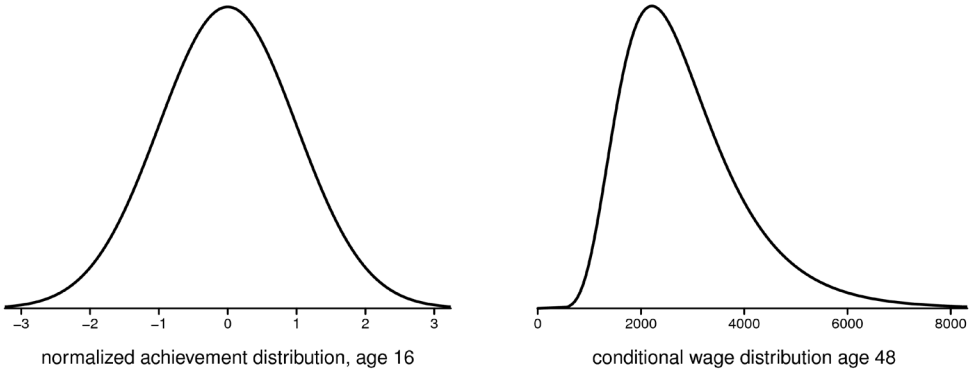

Figure 4: The estimated wage distribution conditional on differences in achievement levels, controlling for parental background. An one percentile in the achievement distribution (left) is associated with an one percentile increase in the wage distribution (right). Data: NCDS 2010.

To keep things simple, 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 . Suppose that the true distribution is lognormal, and given by

$$y \sim lognormal\left(\mu,\sigma^2\right)$$

but that we measure normal data given by

$$y' = \ln(y) \sim normal\left(\mu,\sigma^2\right)$$

In order to catch only the effect of a change in the shape of the distribution, and not the effect of a change in the scale, I will compare the difference of means in the normal distribution with the difference of logged means in the lognormal distribution. This implies that the difference will be expressed in terms of the normalized test scores.

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

$$bias = \left(E\left[y'_t\right] - E\left[y'_0\right]\right) - \left(\ln\left(E\left[y_t\right]\right) - \ln\left(E\left[y_0\right]\right)\right)$$

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

$$bias = \left(\mu_t - \mu_0\right) - \left(\mu_t + 0.5\sigma_t^2 - \mu_0 - 0.5\sigma_0^2\right) = 0.5\left(\sigma_0^2 - \sigma_t^2\right)$$

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 5.

The dependence of qualitative robustness on the variance of the distributions only can be generalized. 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 those distributions.

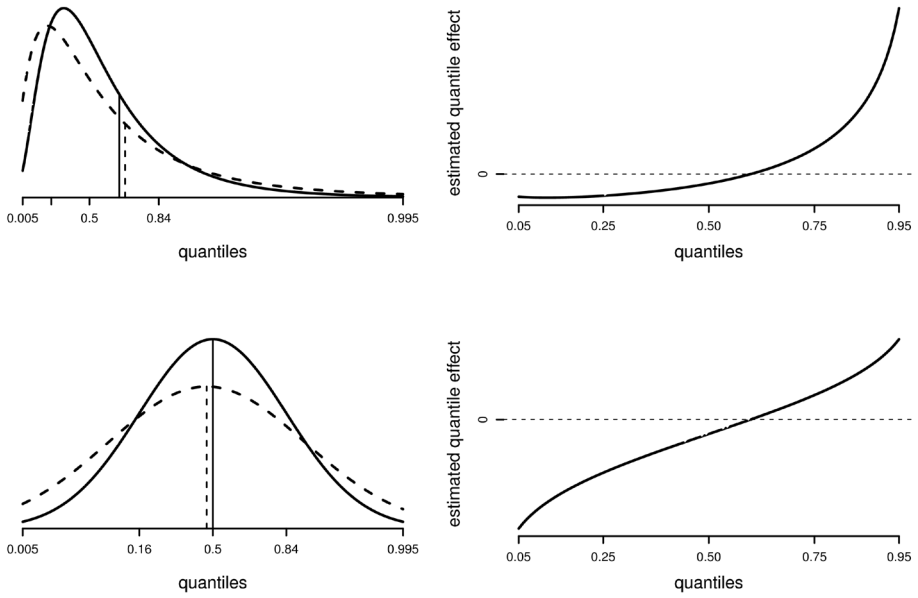

Figure 5: If the true distribution is lognormal (top left panel), normalizing data (bottom left panel) may lead to qualitatively wrong conclusions when comparing distribution means if the group variance differs. In this case, the treatment distribution (dashed lines) has a higher mean in the original data, but appears to have a lower mean after normalization. It should be noted that quantile-based methods (right panels) are qualitatively robust, with a negative effect on all quantiles below about 0.6, and a positive effect on all quantiles above.

The next step is to calibrate this equation by plugging in an empirical σ_0 and σ_t. Let us assume that the width of our control distribution equals that of the reference distribution σ_{ref}, and that the width of the treatment distribution is given by $(1+\beta_\sigma)\sigma_{ref}$:

$$\sigma_0 = \sigma_{ref}$$
$$\sigma_t = (1+\beta_\sigma)\sigma_{ref}$$

Note that , 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.

$$bias = -\sigma_{ref}^2\left(\beta_\sigma + 0.5\beta_\sigma^2\right)$$

How large is the bias in practice? In many cases, variances are more or less constant over treatment, and the bias will be close to zero in accordance with Davison and Sharma (1988). One example where this is clearly not the case is curriculum tracking, the separation of students into different schools or classes based on (estimated) ability. Such stratification almost certainly leads to larger differences between students (cf. Pfeffer 2009, Koerselman (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.

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

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

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 4 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 similar, and are not reported here.

5 Conclusions

The use of mean-based statistical techniques on educational achievement scores is problematic in two ways. On a philosophical level, it is unclear whether statements involving a comparison of score means are statements about the real world, and as such are empirically meaningful. To judge that that this is the case implies that we think score distances to be comparable in different parts of the distribution.

Even if there exists such a true, underlying distribution, we cannot measure it directly at anything above the ordinal level. Instead, we must implicitly or explicitly impose a distribution onto our ordinal data in order to end up with an interval-level score distribution. This causes robustness problems: choosing a different distribution may change our qualitative statements. Groups which have higher score means under one distributional assumption, may have lower average scores under another.

If we insist on using mean-based methods, we can try to get some grips on the size of the robustness problem by imposing different distributional forms on educational test scores, and looking how much estimates differ between the distributions.

I identify the commonly used normal distribution as well as the lognormally distributed conditional wage distribution as two theoretically appealing distributions, and derive an expression for the difference in estimates between the two. It turns out that the difference is quantitatively small, and unlikely to lead to qualitatively different conclusions, even for known inequality-increasing policies like curriculum tracking. In cases where the treatment effect is homogeneous over the distribution, there is no problem at all.

Both the problem of meaningfulness and of robustness can easily be solved by avoiding mean-based methods, and using quantile-based ones such as median regression or quantile regression instead. Even if the bias from using mean-based methods is likely to be small, I advocate using quantile regression as a robustness check, not in the least because the results from such an exercise can be informative. If the effect has the same sign for all quantiles, conclusions are qualitatively robust. If it does not, this an important result in and of itself, and worthy of reporting.

6 References

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

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 coun-tries. 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.

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. Lux-embourg, periodic updating. 2010.

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

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

Fabian Pfeffer. Equality and quality in education. presented at the Youth Inequalities Conference, University College Dublin, Ireland, December 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.