

Swedish Institute for Social Research (SOFI)

Stockholm University

WORKING PAPER 1/2011

**BIAS FROM THE USE OF MEAN-BASED METHODS ON TEST
SCORES**

by

Kristian Koerselman

Bias from the use of mean-based methods on test scores

Kristian Koerselman^{*†‡}

January 3, 2011

Abstract

Economists regularly regress IQ scores or achievement test scores on covariates, for example to evaluate educational policy. These test scores are ordinal measures, and their distributions can take an arbitrary shape, even though they are often constructed to look normal. The ordinality of test scores makes the use of mean-based methods such as OLS is inappropriate: estimates are not robust to changes in test score estimation assumptions and methods. I simulate the magnitude of robustness problems, and show that in practice, problems with mean-based regression of normally distributed test scores are small. Even so, test score distributions with more exotic shapes will need to be transformed before use.

Keywords: *admissible statistics, test scores, educational achievement, item response theory, IQ, PISA.*

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

Economists regularly regress measures of ability or achievement on covariates, for example to evaluate educational policies or teacher performance (cf. Lazear 2003, Todd and Wolpin 2003, Hanushek 2006). I will generally refer to these measures as *test scores*. Examples of test scores range from cognitive and noncognitive skill scores such as IQ scores, to school grades and scores from large international surveys of educational achievement such as PISA or TIMSS. Psychometric theory tells us that mean-based regression of test scores is problematic, something which is often ignored in the economic literature.

Ability and achievement cannot be measured directly, but must be estimated in a two-step process. First, we must devise a test which produces raw scores containing information on the concept we are trying to measure. Then, we must find a suitable function which maps raw scores into the reported test scores. As will become clear further below, the result of this indirect measurement is that we can never measure the underlying trait on a higher level than the ordinal. This has not stopped test makers from reporting score distributions which look cardinal. Often, test makers try to create an approximately normal score distribution, but as can be seen from Figure 1, test scores follow a rather arbitrary distribution at times.

Because of the ordinality of test score data, conclusions drawn from mean-based regression of test scores are not robust to changes in the choice of either the test or the test score mapping. Changing these will give us a similar ranking of students, but expressed in scores with different cardinal values. In many cases, there exists a possible alternative set of values which will lead to a change in regression coefficients large enough to invert their sign.

A question separate from measurement is whether we consider the underlying ability and achievement traits to be ordinal or cardinal. If we think of them as ordinal, we do not only have robustness problems, but problems of

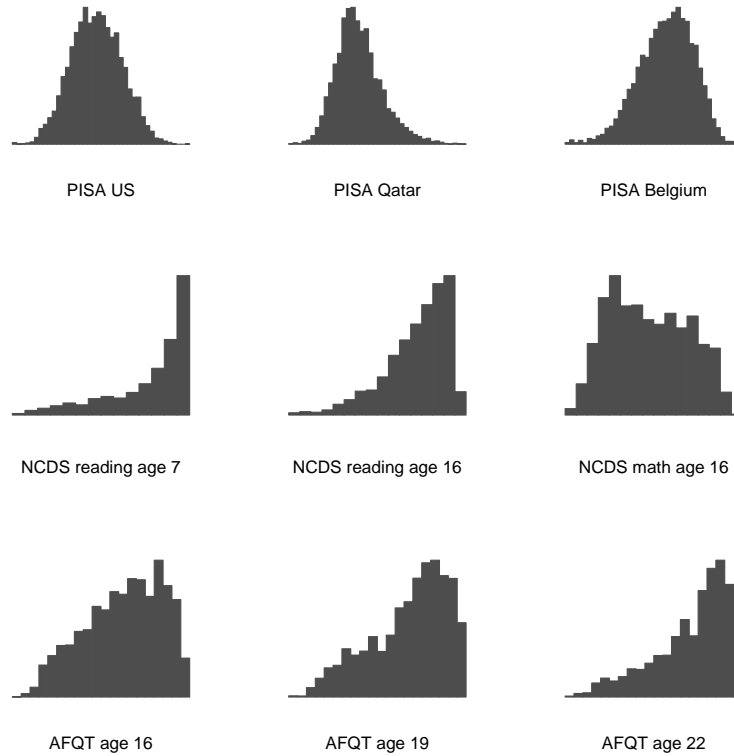


Figure 1: Test score distributions can take different shapes. Top row: weighted country score distributions for PISA 2006 math scores (OECD 2006). Middle row: reading and maths test scores for the NCDS (2010). Bottom row: weighted AFQT scores derived from the ASVAB administered to the NLSY79 (2010) sample.

interpretation as well. Statements comparing averages of ordinal variables are fundamentally meaningless.

In theory, both the problem of qualitative robustness and of meaning can be solved by using methods fitting to ordinal data, such as quantile-based methods. These methods do however add an additional layer of complexity to the models commonly used in educational analysis, and the empiricist may shy away from them in practice. Also, we would like to know how much trust to place in the existing mean-based literature. Is it invalidated by the use of wrong methods?

I argue that as economists we have a clear cardinal interpretation of underlying ability and achievement in the form of human capital. Even if this solves the problem of interpretation, the robustness problem remains because we cannot measure cardinal-level information. However, given that we interpret test scores as a measure of human capital, we can derive bounds on the extent of the problem by making assumptions on what would constitute a reasonable shape of the human capital distribution.

On the basis of theory, we can expect underlying ability and achievement to be either normal or lognormal. The normal is commonly used, but the empirical labor market value of human capital seems to be lognormally distributed. The practical difference in using either distribution when the other is appropriate is small. I illustrate this by calculating bounds for analyzes of curriculum tracking – a policy for which the problems should be relatively grave compared to other educational policies. I find that the variation in the estimate of its effects are an order of magnitude smaller than the point estimates, and that mean-based results are qualitatively robust.

Nevertheless, economists should be aware of the assumptions going into the estimation of test scores. Score distributions with an arbitrary shape, such as the distributions in the middle and bottom rows of Figure 1, should perhaps be transformed into a normal or lognormal distribution if results are to be interpretable and remotely comparable to other studies.

I advocate the use of quantile-based methods when analyzing test scores as a robustness check for mean-based methods. Because the qualitative robustness of mean-based analysis increases in the homogeneity of the estimated effect, the cases in which mean-based comparisons are the least robust are exactly the cases where quantile-based methods will yield the most interesting results, and will be worthy to report in any case.

2 Psychometric theory

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 want our statements to be qualitatively 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 as to which nation is the tallest 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 an ordinal variable are not robust to the choice of scale. Consider *highest completed education*. Using 1 for *primary education*, 2 for *upper secondary education* and 3 for *tertiary education* may or may not give a different ordering of the English and French means compared to using 9 for *tertiary education* instead of 3, even if $(1, 2, 9)$ are just as good a representation of the ordinal levels as are $(1, 2, 3)$.

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.

A related but distinct problem is that of meaningfulness. We calculate statistics on 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 empirically *meaningful* (cf. Hand 2004, section 2.4.1). A statement like “*The mean completed education in England is 1.8.*” makes no

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.

sense because it is not a statement on education in England as much as on the mean of a vector of arbitrary numbers stored on our computer.

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 education in England and France, and conclude that they are significantly *different*: that the English and French samples are not likely 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.

Test scores are used as a measure of a variety of concepts. They are designed to capture variables like intelligence or ability, proficiency at a certain task, or learning. Below, I will refer to the underlying variable as ‘achievement’ even though the reasoning applies to other variables just as well.

Achievement cannot be observed directly, but must be estimated using some kind of framework. These can be divided into two broad categories.

The simpler of the two is called Classical test theory or CTT. In CTT, the test score is a linear transformation of the proportion of test items or questions

answered correctly. This is the kind of scoring we perhaps remember from our own time in school.

CTT is based on a true score model

$$x = t + \varepsilon$$

where t is the true, underlying probability of the student answering questions correctly, and x is the observed proportion of questions actually answered correctly. The error ε arises because the number of questions is limited, adding noise to the estimate. We use x as the estimate of t .

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: they are 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 2). 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)

The difference in the skew of the score distribution affects our estimated mean effects. If we were to compare means between a treatment and a control group on the basis of the hard test, we would weight the right tail of the distribution more heavily, whereas if we were to compare means on the basis of the easy test, we would weight the left tail more.

We can interpret CTT scores on a ratio level when speaking about a specific test. We could for example use a change in achievement test scores to identify that the average probability of answering correctly on a specific level of questions has increased. We cannot, however, generalize the result to the scores obtained by a different achievement test, even if both tests are designed to measure the same dimension of achievement. Also, we cannot make ratio-level statements on the effect on the underlying trait. We cannot conclude that ‘mean achievement’ has increased even if mean test scores have.



Figure 2: 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 an *item response function* which describes the probability of giving a specific response or answer to the specific item or question the function refers to. Often, the response categories are ‘right’ and ‘wrong’, and we get an item response function of the form

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

This function is illustrated in Figure 3.

$\mathbb{P}(y_{ij} = 1)$ is the probability of student i answering question j correctly, θ_i is the level of student achievement we are interested in, and b_j is overall question

difficulty. The inflexion point of the response function lies at $b_j = \theta_i$, and we say that student achievement and question difficulty are equal at this point.

The limiting probability of answering the question correctly for extremely low levels of achievement is given by c_j . It can be interpreted as the probability of answering correctly when making an uninformed guess. The upper probability limit is assumed to be one.

Question discrimination a_j determines the rate at which students get better at answering the question when they have a higher level of achievement. A question which everyone answers equally well has zero discrimination. a_j may even turn negative if answering correctly on question j correlates negatively with answering correctly on the others. Questions with low or negative discrimination are however regularly discarded from test databases.

There are model variations where one or more item parameters are fixed or otherwise restricted, mainly because there is relatively little information contained in each response. When c is set to zero, and a to one, we obtain the 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.

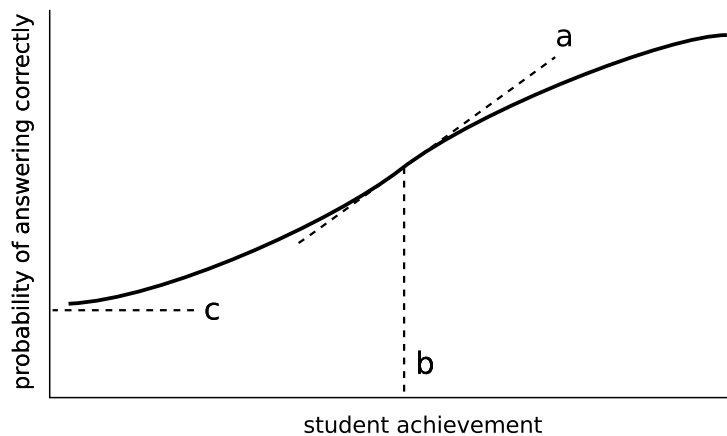


Figure 3: An item response function gives the probability of a student answering a certain question correctly as a function of his achievement. Question parameters a (discrimination), b (difficulty) and c (guessing) are illustrated in the figure.

IRT models remove much of the CTT models' dependence on question difficulty. 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, the distance $\theta_i - b_j$ must be relatively close to zero, just as when he does not do unusually well on questions below his level.

In theory, we could even give some students easy questions and others hard ones, and their results would still be comparable, albeit at the cost of test efficiency. It would be tempting to think that IRT models therefore also solve the CTT models' ordinality problem, but unfortunately it returns in a different guise.

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.

We can also change the higher moments of the distribution by estimating a different functional form. The horizontal achievement and difficulty 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 distribution of θ and b are stretched out in one tail and compressed in the other (Lord 1980, p. 85), lending the score distribution arbitrary skew.

While CTT distances are a product of the particular test taken, IRT distances depend on the equations which we use to map raw scores into test scores. In both cases, we can reasonably change our methods, and obtain a test score which follows a different distribution.

Given what we know about admissible statistics, meaningfulness, and the process by which test scores are estimated, how should we deal with test scores? Regression analysis is the comparison of means conditional on the relevant variables. The sensibility of regressing test scores equals the sensibility 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 be able to accept that there exists an amount of underlying improvement in the lower tail of the distribution which is equivalent to a certain amount of improvement in the higher tail, that the failure of one student to learn a certain skill can be exactly offset by the success of another student at learning a different skill.

Whatever our thoughts on meaningfulness, the problem of robustness however remains. The cardinal values of test scores are arbitrary, and a change in their values will cause a reweighting of observations. For example, in Figure 4, treatment has a positive effect in the right tail, but a negative effect in the left. In the original data (top left panel), the positive effect outweighs the negative, and the mean treatment effect is positive. In the normalized version of the data however (bottom left panel), the right tail is given less weight, and the mean treatment effect turns negative.

As can be seen from the right side of the figure, quantile regression is qualitatively robust in the sense that for each quantile, the estimated effect has the same sign both in the true and in the normalized distribution.

In conclusion, both the problem of meaningfulness and of robustness can be solved by using statistics of the ordinal level on test scores. A statement like ‘the median student in the treatment distribution has higher achievement than the median student in the control distribution’ needs no cardinal interpretation, and quantile-based methods are qualitatively robust. Psychometric theory would discourage us from using mean-based analysis as it is based on information we cannot actually measure.

3 Economic practice

The argument of admissible statistics for refraining from mean-based analysis on test scores is the strongest if underlying achievement is fundamentally

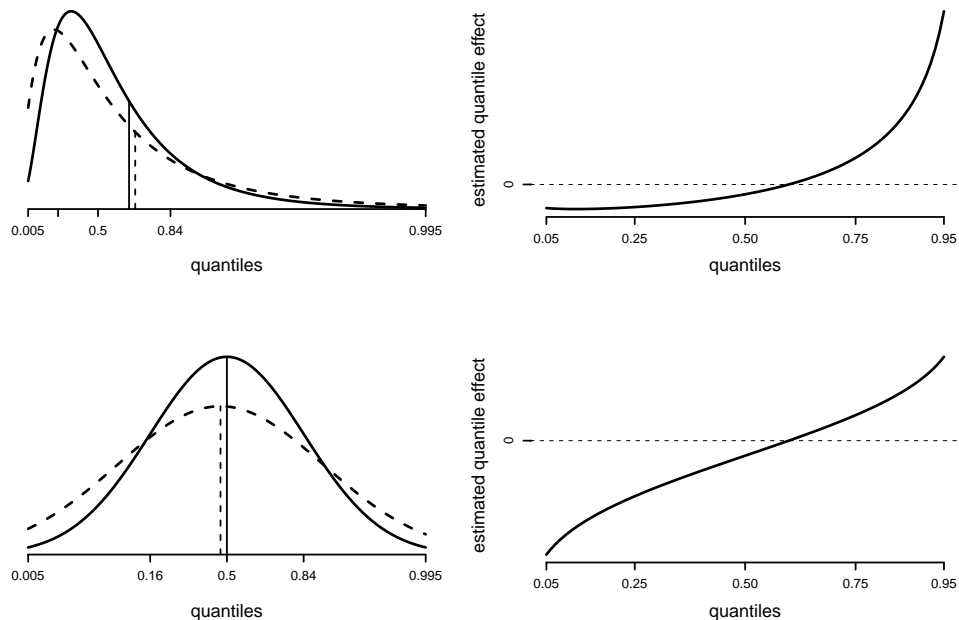


Figure 4: If the true distribution (top left panel) differs from the imposed one (bottom left panel), this may lead to qualitatively wrong conclusions when comparing distribution means. In this case, the treatment distribution (dashed lines) has a higher mean in the original data, but appears to have 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.

ordinal. As economists, we may however feel that it is cardinal because we have a cardinal interpretation of achievement readily available to us: human capital (cf. Hanushek 2006, section 2). While it may seem questionable to state that one student has twice as much ‘achievement’ or ‘intelligence’ compared to another, saying that he has twice as much human capital makes more sense.

Even if the underlying variable is cardinal, its measurement will still be ordinal. We can try to solve this by imposing assumed distributional forms onto our ordinal measurements, and hope that they approximate the underlying reality. If we happen to choose the correct form, all will be well, but the more wrong we are, the more our empirical results will tend to differ from the true relationship.

The use of ordinal methods guards against true distributional forms that are *arbitrarily* different from the assumed distribution, but we may feel that this is overly conservative. If the true distribution comes from a smaller family of possible distributions, the scope for robustness problems will be smaller too.

Which possible underlying distributions of achievement seem reasonable? The normal distribution stands out as a natural candidate. It has theoretical appeal, as it emerges from an addition of many independent draws from an arbitrary, finite distribution per the central limit theorem. If we think of learning as an additive random process, where each day's new learning is a random draw to be added to the existing stock, achievement will be normal.

A second appealing distribution is the the lognormal. It has a similar relationship to the central limit theorem as the normal: if we multiply rather than add the (positive) draws, we will end up with a lognormal distribution. We can think of learning as a process in which students start from the same baseline, and learn at small, random rates each day, finally arriving at their test-day achievement level. The drawing of learning rates rather than amounts implies that we think that past and future learning is correlated on an absolute level, so that higher achieving students are expected to acquire more 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, the achievement distribution will have right skew if high ability individuals also put more effort, time or other resources into learning (cf. Becker 1964/1993, p. 100).

In spite of this, the test scores we use usually either have an arbitrary distribution, or an approximately normal one. Sometimes test makers actively tweak tests to yield a normal distribution, sometimes it arises by chance.

If we think of achievement as human capital, we can estimate the distribution of its value empirically. To do so, I use the UK National Child Development Study (NCDS 2010). I select males that are full-time employed at age 48, and calculate the principal component of their normalized age 11 and 16 test scores.

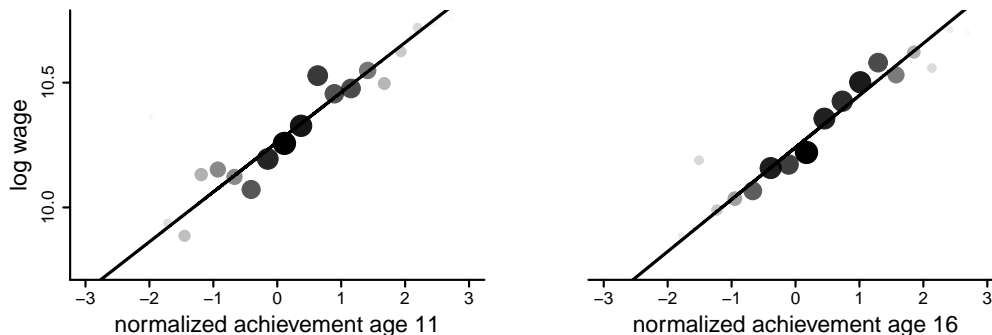


Figure 5: Average logged gross wages of 48-year old full-time employed males for different achievement levels (circles, circle area and color is proportionate to the number of observations) and the regression line through the unaveraged data. Data: NCDS 2010.

Figure 5 shows average logged age 48 gross wages for different achievement intervals at age 11 and 16 (circles). There is an approximately linear relationship between test scores and logged wages, which implies an exponential mapping from scores to wages. This is indeed the standard assumption in the economics of education literature (cf. Lazear 2003, Hanushek 2006).

We can find the slope of the relationship by regressing wages Y on the principal component of test scores y .

$$\ln Y = a + by + \varepsilon$$

The estimated values of b can be found in the first column of Table 2. They are moderately large at wage differences up to 0.21 logs for a one standard deviation increase in age 16 test scores.

In column (2) I add controls for socioeconomic background to the equation above. I arrive at an association between test scores and log wages of 0.17 for the age 11 achievement distribution and 0.18 for age 16.

Depending on what we want to condition the wage distribution on, we can add more controls. For example, we can add age 7 scores as a proxy for ability. The estimates of the effect of later age test scores are then reduced

dependent variable: log wages	(1)	(2)	(3)	(4)
age 11 score	0.20	0.17	0.13	0.06
age 16 score	0.21	0.18	0.14	0.04
parental background		yes	yes	yes
age 7 scores			yes	yes
highest educational attainment				yes

Table 2: Eight separate estimates of the relationship between standardized test scores and logged age 48 wages for full-time employed males in the British 1958 cohort. Data: NCDS 2010.

to 0.13 and 0.14 respectively, as can be seen from column (3). Column (4) shows the estimates when I control for (endogenous) educational attainment as well. This reduces the estimates to 0.06 and 0.04.

Other authors have found estimates in the range of 0.07–0.16 for specifications similar to the first (Lazear 2003), 0.09–0.15 for specifications similar to the second (Murnane et al. 2000), and 0.06–0.16 for the fourth (Murnane et al. 2000, Altonji and Pierret 2001, Galindo-Rueda 2003, Lazear 2003, Speakman and Welch 2006).

In specifications (1) and (2), my estimates are a bit higher than those of the other authors. Robustness checks show that this is because I normalize the test scores before taking their principal components. Another difference in my method is that I use old-age wages, and the effects of the impact of test scores should rise with age (Altonji and Pierret 2001, Galindo-Rueda 2003, Lange 2007). I have tried regressing age 33 wages instead, but this yields roughly similar results in my sample.

The lognormal distribution can have different levels of skew. How skewed would a conditional wage distribution be? We can map the standardized score distribution y into the estimated conditional wage distribution \hat{Y} according to

$$\hat{Y} = \exp(\hat{a} + \hat{b}y)$$

The conditional wages \hat{Y} will be distributed lognormally with the distribution parameter σ equal to \hat{b} . The larger σ , the larger the skew of the lognormal. For low values of σ , the conditional wage distribution will look like the nor-

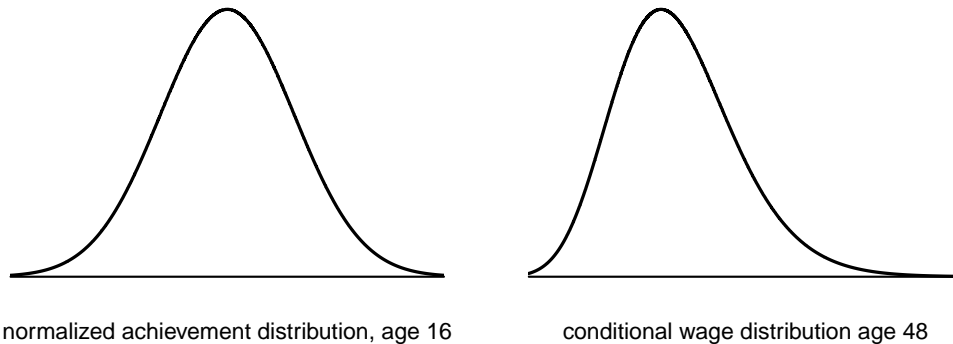


Figure 6: The estimated wage distribution conditional on differences in achievement levels, controlling for parental background. A one percentile in the achievement distribution (left) is associated with a one percentile increase in the wage distribution (right). Data: NCDS 2010.

mal. The second column of Table 2 gives the largest values of b which we still might reasonably call causal at 0.18, even though the true causal effect is probably much smaller. The lognormal distribution we obtain by setting σ equal to 0.18 can be seen from Figure 6.

Suppose that the true cardinal achievement distribution is given by the conditional wage distribution, but that we use a normal test score distribution for our analysis. It is possible to calculate the bias arising in OLS estimates because of this.

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 β_μ . Suppose that the true distribution is lognormal, and given by

$$Y \sim \text{LN}(\mu, \sigma^2),$$

but that we measure normal data given by

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

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:

$$\text{bias} = (\mathbb{E}[y_t] - \mathbb{E}[y_0]) - (\ln(\mathbb{E}[Y_t]) - \ln(\mathbb{E}[Y_0])).$$

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

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

Let us define β_σ as the amount by which the treatment distribution is wider than the control distribution. Note that β_σ is expressed in control group standard deviations.

$$\sigma_t = (1 + \beta_\sigma)\sigma_0$$

Rewriting the earlier equation in terms of σ_0 and β_σ we then get

$$\text{bias} = -\sigma_0^2 \left(\beta_\sigma + \frac{1}{2}\beta_\sigma^2 \right).$$

This shows that the amount of bias generated by assuming a normal distribution where the lognormal distribution is appropriate is independent of the treatment effect on the mean, but dependent on the difference in variance between treatment and control groups. A relatively larger variance in the treatment group will lead to a negative bias in the estimate of the treatment effect, and vice versa.

The dependence of qualitative robustness on the higher moments 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.

	(1)	(2)	(3)	(4)	(5)
Paper	β_μ	β_σ	σ_0	bias	corrected β_μ
Hanushek and Woessmann (2006)	-0.179	0.101	0.182	-0.004	-0.175
Pekkarinen et al. (2009)	-0.007	0.009	0.182	0.000	-0.007
Duflo et al. (2008)	0.175	0.042	0.182	-0.001	0.176

Table 3: Estimated treatment effects of curriculum from a number of selected papers. The last column shows the estimated effect under the assumed lognormal distribution.

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 (Pfeffer 2009, Koerselman (forthcoming)). We can thus use curriculum tracking as a kind of worst-case scenario for educational policy analysis.

I have selected three empirical papers from the literature on the subject from which to get empirical values for β_σ . 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 column in Table 3 shows standardized estimated treatment effects on the mean from these papers. The second column contains the standardized effects on the distributions' standard deviations. The third column contains the parameter σ_0 , which determines the skew of the assumed underlying human capital distribution. I assume the human capital distribution to have a σ equal to 0.18.

The fourth column contains the size of the bias, and the fifth the updated mean treatment effect. The size of the bias is quantitatively small; far below 0.01 of a standard deviation in test scores for all three papers. This is not enough to change the papers' respective qualitative conclusions.

The estimate \hat{b} underlying the conditional wage distribution used for this analysis is probably an overestimate of the causal relationship between test scores and wages. In the estimate, we omit both the effect of early age ability and of differences in later age educational attainment. Were we to use a smaller value of \hat{b} , the corresponding biases would be smaller in size as well.

4 Discussion

Test scores contain little information on a higher level than the ordinal. We can rank students, but the cardinal values attached to those ranks are essentially arbitrary. The ordinality of test scores makes conclusions based on a comparison of means sensitive to the exact method of producing those scores. In theory, regression coefficients may even change sign when we give students a different test, or change the mapping of raw scores into test scores.

It is a matter of opinion whether there exists an underlying cardinal concept of ability or achievement. Human capital is however much more straightforward to interpret cardinally. Stating that "*Individual A has twice as much human capital as individual B.*" is more sensible than stating that he is twice as intelligent.

If we interpret test scores as a measure of human capital, we may impose an assumed or estimated distributional shape of human capital on the ordinal scores. The better we succeed in doing so, the truer our mean-based conclusions will be.

When scores are made to follow either a normal or a lognormal distribution of limited skew, we can reasonably think of them approximating the underlying human capital distribution. The difference between regression coefficients

based on either distribution is numerically small in most cases, and the reader may conclude that to limit oneself to the the use of ordinal methods on test scores is overly prudent.

At other times however, test score distributions seem to have an arbitrary shape. If we use mean-based methods on those distributions, we must explicitly consider whether the empirical distribution fits the distribution we assume the underlying variable to have. There are cases where the test was designed to yield normal scores in a larger population, but where we only observe a subsample. In such cases, the scores may contain true cardinal information. In other cases the distribution may be truly arbitrary, and it may be a good idea to transform it into a normal or lognormal, or to use percentile scores instead.

I advocate the use of methods such as quantile regression on test scores, at the very least as a kind of robustness check. The bias produced by using the wrong distribution of test scores increases in the heterogeneity of the estimated effect. Those cases where mean-based analysis is the least robust will thus also be the cases where quantile regression will give the most interesting results, worthy of reporting in and of themselves.

Acknowledgments

I thank Markus Jääntti, Denny Borsboom, René Geerling, Annemarie Zand Scholten, Christer Gerdes and Anders Stenberg 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.G. Altonji and C.R. Pierret. Employer Learning and Statistical Discrimination. *Quarterly Journal of Economics*, 116(1):313–350, 2001. ISSN

0033-5533.

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.

F. Galindo-Rueda. Employer Learning and Schooling-Related Statistical Discrimination in Britain. IZA DP 778, 2003.

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

E.A. Hanushek. School resources. *Handbook of the Economics of Education*, 2:865–908, 2006. ISSN 1574-0692.

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.

F. Lange. The speed of employer learning. *Journal of Labor Economics*, 25(1):1–35, 2007.

E.P. Lazear. Teacher incentives. *Swedish Economic Policy Review*, 10(2):179–214, 2003. ISSN 1400-1829.

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.

- R.J. Murnane, J.B. Willett, Y. Duhaldeborde, and J.H. Tyler. How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, 19(4):547–568, 2000. ISSN 1520-6688.
- 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.
- Fabian Pfeffer. Equality and quality in education. presented at the Youth Inequalities Conference, University College Dublin, Ireland, December 2009.
- R. Speakman and F. Welch. Using wages to infer school quality. *Handbook of the Economics of Education*, 2:813–864, 2006. ISSN 1574-0692.
- Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.
- Petra Todd and Kenneth Wolpin. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113:F3–F33, 2003.
- U.S. Bureau of Labor Statistics. National Longitudinal Study of Youth 79. 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.