SOCIAL FORSKNINGS INSTITUTTET

James McIntosh
Martin D. Munk

**20:2007** WORKING PAPER

# WHAT DO EDUCATIONAL TEST SCORES REALLY MEASURE?

RESEARCH DEPARTMENT OF CHILDREN, INTEGRATION AND EQUAL OPPORTUNITY

# *WHAT DO EDUCATIONAL TEST SCORES REALLY MEASURE?*

***James McIntosh***
***Martin D. Munk***

***Working Paper*** ***20:2007***

# WHAT DO EDUCATIONAL TEST SCORES REALLY MEASURE?

**James McIntosh**[*,†] **and Martin D. Munk**

[*]Economics Department
Concordia University
1455 De Maisonneuve Blvd. W.
Montreal Quebec, H3G 1M8, Canada.

[†]Danish National Institute of Social Research
Herluf Trolles Gade 11
DK-1052 Copenhagen K, Denmark

June 29, 2007

**E-mail Addresses and Telephone Numbers:** jamesm@vax2.concordia.ca 001 514 848 2424 Ex.3910. mdm@sfi.dk 45 33 69 77 10.
**Keywords:** Human Capital, Educational Production Functions, Test Scores, Ability, and Unobservable Heterogeneity.
**JEL Classification Numbers:** I21, C25.

**Abstract**

Latent class Poisson count models are used to analyze a sample of Danish test score results from a cohort of individuals born in 1954-55 and tested in 1968. The procedure takes account of unobservable effects as well as excessive zeros in the data. The bulk of unobservable effects are uncorrelated with observable parental attributes and, thus, are environmental rather than genetic in origin. We show that the test scores measure manifest or measured ability as it has evolved over the life of the respondent and is, thus, more a product of the human capital formation process than some latent or fundamental measure of pure cognitive ability. We find that variables which are not closely associated with traditional notions of intelligence explain a significant proportion of the variation in test scores. This adds to the complexity of interpreting test scores and suggests that school culture, attitudes, and possible incentive problems make it more difficult to elicit true values of what the tests measure.

## 1 Introduction

Educational testing plays a very important role in our society. Educators use early test score results to determine the most appropriate type of education stream that an individual should follow. At the social policy level the relation between test scores and the individual's socioeconomic characteristics is used to

inform decision makers about the need for interventions to assist disadvantaged groups or to determine how much should be spent on the educational systems and at what level. There are two reasons for this. First, educational tests taken at fairly young ages have been shown to be good predictors of later educational attainments as well as success in the labour market. Secondly, it is believed in some quarters that ability or being smart is what really counts and educational test score results accurately reveal cognitive ability or innate intelligence.

On the second point, however, there are serious differences of opinion. Herrnstein and Murray (1995 p. 22) refer to educational test scores as IQ tests and write

> "IQ scores match, to a first degree, whatever people mean when they use the word *intelligent* or *smart* in ordinary language".

On the other hand Brody (1992 p. x) says

> "I think that individual differences in intelligence, as assessed by standardized tests, relate to what individuals learn in school and to their social mobility. And I think that scores on such tests are related, albeit weakly related, to race and social class background".

Such divergent views reveal a serious issue which needs to be addressed. What do test scores tell us about individual ability? Knowing exactly what test scores measure is particularly important in determining schooling options. If, for example, it is mistakenly believed that test scores primarily measure intelligence then individuals who do poorly on these tests could be labeled as being intellectually challenged and encouraged or forced into vocational or less academically oriented educational alternatives. Or what is even more damaging to them, they could be prevented from participating in programmes which address the problems that lead to their poor test score performance. On the other hand, if test score results also reflect the stock of the child's human capital then education policy might be more usefully focused on remedial programmes that help

disadvantaged students overcome the problems that are caused by low parental investments.

One of the purposes of this research is first to review the literature on test score determinants to see what it has to say on the issue. Although the question that we pose is not often considered explicitly[1], there is a considerable amount of information that is relevant and revealing of the content of test score outcomes. However, the main objective of the project is to analyze a set of test scores that were obtained from a representative sample of Danish students aged fourteen in 1968. This data comes from The Longitudinal Survey of Youth and is described in Hansen (1995) and McIntosh and Munk (2007). The data set contains information on test scores as well as some information on the features of the households in which the respondents lived in 1968 together with some of the characteristics of their parents.

The test score results consist of the number correct answers to the test which we analyze using count models. Our preferred specification is the Latent Class Poisson Model of Wedel *et al* (1993) which takes account of the presence of excessive zeros as well as dealing directly with the problem unobservable respondent characteristics.

To briefly summarize our results, like many other studies, we find that test scores depend significantly on the characteristics of the households in which the respondents lived at age fourteen as well as the characteristics of their parents. But we also found that variables like attitudes to school and the scholastic abilities of the respondent's classmates, variables which have nothing to do with how smart the individual is, were also highly significant covariates in test score equations. As a result we identify a new problem which relates to whether testing procedures can actually elicit the true value of what test scores measure. This leads to a rather different perspective on the meaning of test score results.

From our review of papers using the value added approach we concluded from the low values of the coefficients of the lagged test score that test scores evolve. This process is dynamic and depends on the individual's capacity to accumulate

[1] Neal and Johnson (1996 p. 890) are among the first to make this point.

various types of human capital. If, as Herrnstein and Murray believe, there is a measure which is 'substantially heritable' and invariant over the individual's life this is not what test scores measure.

The rest of the paper is organized as follows: section 2 reviews test score analysis literature. Section 3 describes the data used in the analysis. Section 4 gives a detailed description of the statistical models used, section 5 develops results and discusses their implications.

# 2 The Test Score Literature

We begin this short review of the literature by summarizing two formal models of test score determination. Todd and Wolpin (2003, 2004) examine test scores in a production framework first suggested by Ben Porath (1967). Here achievements or test scores are related to the histories of two input vectors by assuming

$$T_i(a_i) = T[Y_i^c(a_i), Y_i^e(a_i), \mu_i(0), \epsilon_i(a_i)] \tag{1}$$

In equation (1) $T_i(a_i)$ is the test score of individual $i$ at age $a$ , The two $Y$ variables are histories of input vectors up to age $a_i$. The first are chosen by the parents and the second consist of exogenous inputs, hence the superscripts $c$ and $e$. These investments which are made in the child as it develops contribute the child's stock of human capital. $\mu_i(0)$ is what they refer to as 'the child's endowed mental capacity or ("ability")' and $\epsilon_i(a_i)$ is a measurement error.

Hansen *et al* (2004) propose a model which is somewhat different in structure. Their test score equation is

$$T_i(s_i) = \mu(s_i) + \lambda(s_i)f_i + \varepsilon_i(s_i) \tag{2}$$

Unlike equation (1) the focus of attention in equation (2) is on the number of years of schooling attained when the score was administered, rather than age. They refer to $f_i$ as latent ability, or fundamental cognitive ability, or just IQ, whereas the test score is manifest ability and is a measure of observed achievement. In this formulation achievement or manifest ability, as measured

by the test score, is determined by a zero mean IQ variable mediated by a scaling factor which depends on years of school attained together with a mean which depends on schooling as well as individual covariates. In both models there is additional complexity since the $Y^c$ variables in the Todd-Wolpin model and the level of schooling in Hansen *et al* model depend on the child's ability, a variable which is not observed by the researcher. A major consequence of this latter feature of the processes by which test score results are generated is that unobservable variables, the most important of which is latent ability or IQ, play a key role. These are likely to be correlated with the variables that are usually included as explanatory variables in regressions which explain test score results so that failing to take account of these unobservables will lead to misleading statistical results.

The Todd-Wolpin model highlights another important feature of the test score production process and that is the intertemporal dimension of the input structure. The timing of inputs plays a crucial role in their model because they recognize that test score results are determined by a capital accumulation process which involves the whole history of investments that parents make in their children. They also note that the estimation of such models is particularly difficult even when there is more than one age at which test scores were obtained. When there are test scores available at different ages but no information on parental investments prior to the last test score, geometric distributed lag models can be used to estimate the effects of the unobserved parental contributions which are captured by the inclusion of the lagged test score. These are referred to in the literature as 'value added models'. However, this procedure is less satisfactory than it first appears because of the moving average error that is induced by replacing the lagged parental inputs with the lagged value of the test score and the untreated correlation between the regressors and the unobservable ability variable. It should also be pointed that these distributed lag models are problematic in terms of their underlying assumptions. The importance of input effects depend on the size of the lag associated with the effect rather than the more plausible assumption that it should be the age when the investment oc-

curred. Using the US National Longitudinal Survey of Youth data they actually reject these models in favour of models based on the observed lagged values of parental inputs and which control for unobservable child effects so this criticism has empirical support as well.

Both models share two additional features. First, test scores are determined in a separable way by environmental variables or child specific investments and a measure of natural or innate ability. These ability measures, $\mu_i(0)$ and $f_i$, do not depend on any of the parental variables so that inheritance mechanisms play no role in the determination of cognitive ability. While there is some debate over the extent to which IQ is inherited (see Daniels *et al* (1997) for a survey of the issues involved) the position that none of it is inherited is highly implausible. Secondly, ability is seen as unidimensional. There is much evidence against this hypothesis as noted by Brody (1992) and Hunt (1997). However, as Todd and Wolpin (2003) point out, this is just a simplifying assumption and these models easily generalize to accommodate more complex representations of ability.

Todd and Wolpin also examine the data requirements for the implementation of these types of models and note the substantial difficulties that arise and the assumptions that have to be made when inputs are observable at only one point of time. What is missing here, however, is that quite often there is no information on any type of input at all. Instead, surveys or registers provide information on the characteristics of the respondent's parents like occupation and final level of education attained as well as some information on the household in which the respondent grew up like household income, parental employment experience or participation in welfare programmes. Less often there is information on what household conditions were like for the respondent in terms of whether he or she was read to as a child, got help with homework, was praised when successful, was exposed to music and cultural events etc.

When there are only family or household characteristics available it is very difficult to relate these to the actual production process of test score attainment. Do respondents with well educated parents do well because their parents invested more resources in them, or because their parents were better role mod-

els, or because they provided them with better genetic endowments? These questions have no answers when this is the only data available.

Most of the research which examines test score performance relies on data which is fairly limited in scope. Todd and Wolpin (2004) found that having observable investments as well as the ages at which they were made mattered as far as the results were concerned. Consequently, some caution should be exercised in interpreting the results that researchers have found when such data limitations are present. We now turn to this literature.

Feinstein and Symons (1999) apply a value added model to the British National Development Survey which contains all children born in Britain between March 3 and 9, 1958. Their preferred test score is a mathematics score taken at ages eleven and sixteen; the first score is used as the lagged dependent variable. An instrumental procedure is used to deal with the possible dependence of the lagged test score on unobservables. Although there is no information on inputs, current or lagged, there is some information on household activities and the characteristics of the school which the respondent attended. In addition, the data base is a cohort so that all the respondents are the same age, took the test score examinations at the same time, and aside from local variation experienced the same social and economic conditions. Schooling experiences can differ because of different choices made at earlier ages but the authors control for potential endogeneity problems here as well.

They find that parental interest and peer group variables to be the most significant covariates with family background variables like parental education and socioeconomic status, and the number of siblings playing a significant but less important role. Their peer group variables are the socioeconomic characteristics of the respondent's classmates. They also include the type of school or stream of the respondent. Going to a grammar school turns out to be the best alternative in terms of getting the highest mathematics score at age sixteen. This result might be seen as a consequence of a selection process whereby the smartest and most able students go to a grammar school. This is, in fact, not the case because the age eleven test score controls for part of that and the similarity of the

ordinary least squares and instrumental variables estimators suggest that the result is not due to unobservables. What is actually happening is that going to a grammar school whose curriculum includes a major mathematics component improves the test score in mathematics. This is just confirmation of the results of Winship and Korenman (1997 p. 231) or Hansen et al (2004 p. 83) that 'staying in school makes you smarter'.

Their value added model also provides additional insight into the true nature of test scores. They obtain an instrumental variable estimate of the coefficient of the age eleven test score variable of 0.54. This suggests to us that measured ability is dynamic and evolves over time and reflects the effects of schooling and continuing parental investments. Neal and Johnson (1996, Table A3) come to the same conclusion by comparing black-white differences measured at different test score ages.

In another value added study Segal (2005) using the American National Educational Longitudinal Survey finds that including a variable which indicates how well behaved the respondent was in grade eight explains a significant proportion of the grade ten test score even when the grade eight score is included. The only other variables that matter are having a poorly educated father and coming from a broken home. While this is an interesting result some care should be taken in its interpretation since no correction was made for the possible correlation between the lagged test score and unobservables. In addition the behaviour score could be highly correlated with the other household and family background variables used in the study.

There are a large number of studies that use simple statistical methods and regress raw or normalized test scores on various sets of covariates. Since these make no attempt to deal with unobservables it is our view that less weight should be placed on their results because of the possible biases in the estimated coefficients.

Fryer and Levitt (2004) analyze a sample of children whose average age is 66 months using the American Early Childhood Longitudinal Study. The large sample size together with the wealth of information that is available for

each respondent make it somewhat unusual. However, the results obtained are typical of what most researchers find. Family background variables like parents socioeconomic status (education and occupation), home characteristics like books at home, being read to, parents being welfare recipients, maternal characteristics like being a teenage mother etc. all turn out to be significant. There are over one hundred regressors of which approximately thirty percent are significant.

Other studies in chronological order are Zajonc and Markus (1975), Gordon (1976), Scarr and Weinberg (1978), Eckland (1979), Paulhus and Shaffer (1981), Steelman *et al* (1983), Neal and Johnson (1996), Peters and Mullis (1997), and Albernaz *et al* (2002).

# 3 Data and Summary Statistics

The data comes from a survey carried out by The Danish National Institute of Social Research and initiated by E.J. Hansen (1995). Information on 3151 subjects was collected during the period 1968/1969. The data includes information about the subject's attitude towards school (i.e. whether the respondent likes or dislikes school), taxable family income, the employment status of the subject's mother (i.e. whether or not she remained at home during his/her childhood), the presence of financial problems for the family, marriage status of the subject's parents, the number of siblings, the teacher's evaluation of the school class in which the respondent belonged when the tests were taken, the subject's living conditions as a child, the occupations and educational attainments of the subject's parents, and the subject's test scores. These tests were conducted when the subjects were 13-15 years old in 1968.

There are three components of the test, a verbal test, a spatial test, and an inductive reasoning test. The first two turned out to be unsuitable for our purposes so we focussed exclusively on the third test score[2]. The test scores are

[2] The verbal reasoning test actually contained some questions of a mathematical nature so it was not a pure verbal test. Tests which examine different dimensions of ability simultaneously are difficult to analyze and produce results which are even more difficult to interpret. The spatial test failed to explain any of the variation in final educational attainments described in

based on the number of correct answers obtained in each test. Table 1 contains the relevant summary statistics for the third test score.

Among the subjects there were 17 people (0.54% of the sample) whose IQ scores were not observed; these subjects was not included. There are also 113 records indicating test scores of zero for all three tests. This is an interesting characteristic of the sample. As we explain later, conventional statistical models fail to account for their presence. Moreover, the fact that there are more zeros than would have occurred purely by chance suggests that there may be problems associated with getting respondents to truthfully convey their responses to the tests.

Table 2 contains information on the variables which are used to explain the test scores. In the group headed by School Variables the attitudes to school are dummy variables with the obvious interpretation. The variable school class quality represents the teachers opinion of the average academic ability of the class. This is a dummy variable which takes the value one if the class was very good or excellent. The household variables have a straightforward interpretation. Income is household income in thousands of Kroner per month. Mother home means the mother of the respondent spent most of her time at home and did not have a full time job. Broken home means that the respondent did not live with both parents at age fourteen-fifteen. Respondents were also asked whether their parents had financial problems and whether they lived in a large urban center.

Father's occupation is grouped into three categories which correspond to managerial and professional occupations, skilled white and blue collar occupations, and unskilled occupations. For fathers, the education variable is a dummy variables indicating some form of advanced education like a university degree. For mothers the education variable is an indicator of educational qualifications past nine or ten years of school. More detailed categories on parent occupations and education levels were used initially but these were not informative so more aggregated categories were employed.

our paper McIntosh and Munk (2007) so it was also excluded.

# 4 Statistical Models

The most popular way of dealing with test score data has been the use of ordinary least squares. For our data this procedure is not particularly appropriate. As a first step we applied ordinary least squares using robust standard errors to deal with the potential heteroscedasticity arising from the count nature of the data. This procedure does not address the problems of excessive zeros. Table 1 shows that each of the test scores has nearly a four percent zero response. The respondents with these zeros actually took the tests but did not get any correct answers. Although these percentages are quite small regression models do not predict the tails of the distributions very well. The count feature of the data was first addressed by fitting Poisson models. However the test score data is over-dispersed relative to the Poisson model and no account is taken of unobserved factors with this distribution. Negative binomial models were also fitted to the data. In this type of model unobserved heterogeneity is assumed to have a gamma distribution. Conditional on the random effect each test score is assumed to have a Poisson distribution. Integrating out the unobservable effect generates the negative binomial model. This modelling procedure deals with the over-dispersion in the data but like the regression procedures it fails to deal with the excessive number of zeros. It is also less than a completely satisfactory way of dealing with unobserved heterogeneity since the random effect can not be correlated with any of the covariates.

Our procedure, which we now outline, focuses directly on the problems of unobserved heterogeneity and excessive zeros in a count model framework. Our model is a generalization of Heckman and Singer (1984) and belongs the latent class models developed by Wedel *et al* (1993). We assume that there are a finite number of types each with a different level of latent ability. Type $\ell$ has latent ability level which depends on parental characteristics, $X_i^P$. This is our way of allowing individuals to inherit some of their ability from their parents. The expected test score for an individual, $i$, who is of this type, is

$$E(T_{i\ell}) = \exp[X_i\beta_\ell + f_\ell(X_i^P)] = \mu_{i\ell} \tag{3}$$

where $f_\ell(X_i^P)$ is the level of latent ability as it effects the score and $X_i$ is a vector of covariates for individual $i$ which contains $X_i^P$ as a subset. These serve as our somewhat imperfect proxies for human capital.

For each type, all test scores are assumed to have Poisson distributions so that the probability mass function for the test score for a type $\ell$ person is

$$\phi_{i\ell}(x) = \exp(-\mu_{i\ell})\mu_{i\ell}^x/x! \qquad x = 0, 1, 2... \tag{4}$$

If the probability of being type $\ell$ is $p_\ell$ the sample log-likelihood function is

$$\ln(L) = \sum_{i=1}^{N} \ln[\sum_{\ell=1}^{\mathcal{L}} p_\ell \phi_{i\ell}(x_i)] \tag{5}$$

where $\mathcal{L}$ and $N$ are the number of types and the sample size, respectively. The mean and variance of these mixtures of Poisson probability mass functions are

$$\mu_i = \sum_{\ell=1}^{\mathcal{L}} p_\ell \mu_{i\ell} \tag{6}$$

and

$$\sigma_i^2 = \mu_i + \sum_{\ell=1}^{\mathcal{L}} p_\ell (\mu_i - \mu_{i\ell})^2 \tag{7}$$

respectively. The variance is clearly greater than the mean so these mixture distributions are suitable for analyzing data which exhibits overdispersion.

To estimate models of this sort some additional assumptions have to be made. First we assume that

$$f_\ell(X_i^P) = X_i^P \gamma_\ell \tag{8}$$

and write

$$X_i \alpha_\ell = X_i \beta_\ell + X_i^P \gamma_\ell \tag{9}$$

since only $\alpha_\ell = \beta_\ell + (0, \gamma_\ell')'$ can be identified.

The choice of the number of mixtures to apply is an empirical issue to be determined by criteria involving the value of the maximized likelihood together with the number of parameters. The appropriate model to be selected is determined by the data in Table 3. The first line for each test score contains the value of the maximized likelihood function using a single Poisson distribution

with no covariates except a constant term. This serves as a baseline which can be used to compare other models and to construct a pseudo-$R^2$ for each model. Additional mixtures were added until there was no significant increase in the penalized likelihood function or until convergence difficulties were encountered.

# 5 Estimation Results, Discussion, and Conclusions

Parameter estimates appear in Tables 4 and 5 by gender and type. It is clear from Table 3 that the likelihood function continues to increase significantly as the number of mixtures increases. The increase in the number of parameters as the number of mixtures increases is fourteen per mixture. Models with more than three mixtures failed to converge properly and appear not to be identified; hence, the estimated parameters in Tables 4 and 5 involve only three mixtures.

Gender is important. Pooling the two genders together was never supported by a likelihood ratio test as an alternative to separate models for each gender. The value of the ln-likelihood for the pooled sample with a gender dummy is -10072.553 whereas the sum of the ln-likelihoods is -11033.782. This gives a $\chi^2_{45}$ statistic of 77.542 which is large enough to reject the hypothesis that the two genders are the same.

Goodness of fit statistics for all of the estimated models appear in table 7. The data is bimodal for each gender. The first mode occurs at zero. Using mixtures allows all of the models to fit the data very well. The zeros, the first two moments and the actual distribution are well predicted by all of the models. Latent class mixture model have been used by Deb and Trivedi (1997) to deal with the problem of excessive zeros. It is also possible to use more traditional procedures involving the zero inflated model of Lambert (1992) or the hurdle model of Mullahey (1986) but these do not deal with unobservables. The psuedo $R^{2's}$ are 25.9 and 29.1 for boys and girls, respectively which are quite high for sample survey data.

In Tables 4 and 5 the estimated coefficients for the three mixture models are

displayed. There is a set of coefficients for each type. For comparison purposes the results for the unmixed Poisson model are displayed in the last column of the two tables. The most interesting and important feature of our results is the difference across the three types. Type I individuals do well on the test. They have the highest mean and the conditional probability of being Type I given that the respondent scored zero is zero. While their scores depend attitudes to school, whether their class was a good one, the number of siblings and the father having a high level job the coefficients are much smaller than the coefficients for the other two types. It is the constant term which is the most important contributor to the high mean score.

Type III individuals, on the other hand, do poorly on the test and almost all (96% for boys) of the zero scores can be allocated to this type. The respondents in this group appear to be severely disaffected. Disliking school has a catastrophic effect on the score for girls. Type III boys and girls respond negatively to their parents' income and the number of siblings they have. They appear to resent being in a school class where performance levels are high but benefit from a mother who does not work. Type II respondents are an intermediate case. They are more sensitive the their family backgrounds but have a lower constant term indicating a lesser importance of external effects on their performance.

The importance of this result is that it shows that individuals with the same parental types can respond differently to their environments. This is not surprising. Parents with the same characteristics can have radically different abilities and they can provide different types of advantage or disadvantage for their children. Children also have different attitudes and personalities. Affluent households often produce academically successful children but they can also produce problem children and occasionally juvenile delinquents.

That the most successful respondents should be the least dependent on the observable characteristics of the household in which they grew up is a most unusual finding. This is caused by the relatively large value of the intercept term. How is this to be interpreted? One possibility is that the intercept terms

are picking up that part of ability or intelligence which is not inherited from the parents (Type I could just be smarter than the other two types). It could also represent characteristics which are external to the family such as school, neighbourhood, or peer group quality or some of the non-cognitive unobservable benefits that the respondents get from the household in which they resided as children and adolescents. Under this interpretation it is sociocultural, economic, environmental and random factors that are the main drivers of high test score performance rather that inherited genetic factors.

This is an unusual result and we have not encountered anything like it in the economics of education or the behavioural genetics literature. Economists who examine surveys involving adopted children, Björklund et al (2006), Plug and Vijverberg (2005) and Sacerdote (2004), find a larger role for biological than environmental factors. Results obtained by behavioural geneticists like Plomin et al (1997, p 444)[3] lead to even stronger claims. They write:

> "Correlations between adoptive parents and their adoptive children provide a direct estimate of the variance of cognitive abilities accounted for by environmental transmission from parent to child. The near-zero correlations indicate that this environmental component of variance is negligible."

However, the position taken by the behavioural geneticists should be viewed with considerable caution because correlations between parent and child test scores are uninformative about the relation of other variables to child test scores if these variables are uncorrelated with the observable characteristics or attributes of the parents. For example, children living in households which experienced marital difficulties at the time the children took the test could have been adversely affected. To suggest that this could not happen on the basis of near-zero correlations between test scores is, of course, absurd. As we have already indicated, it is the importance of unobservable ability components that are not correlated with parental characteristics that are the issue here. Un-

[3] See also DeFries *et al* (1994),

fortunately, the literature referred to above is not very informative about our result.

The presence of zero test scores raises an interesting issue For us it is highly implausible to believe that a recorded score of zero actually reflects the ability of the respondent getting the zero since it is almost impossible to get all 150 answers wrong even if respondents had randomly selected the answers to the test score questions.[4] Some other process must be at work here. We suspect that the respondents who obtained the zeros were simply unwilling to answer the questions and handed in blank questionnaires. Why they should do this is not clear. There is considerable amount of effort required to get good results on these tests and perhaps not everyone felt obliged to provide that. Refusing to make any effort at all is extreme but the problem of incentives is one which should be considered not just for the zeros but for all of the respondents. In this case the respondents were selected to participate in a research project. They were not asked whether they were willing to participate and nothing depended on their test score results so they had no incentive to produce their best possible results.

Of course, perceived benefits and costs are not the only things that matter in determining how much effort should put into answering test score questions. Attitudes and class-room culture are also important and these may explain why liking school[5] and the quality of the class were among the more important regressors in all of the models. It could be argued that not liking school is just a way that low ability individuals rationalize their failings. This is not likely to be the case, however, because even amongst the zeros a large majority of the respondents said that they liked school. It is quite reasonable to find that

[4] The impact of test scores on attainments is analyzed in a companion paper McIntosh and Munk (2007). In that paper we showed that test score performance was an important variable in the explanation of attainments. This is not surprising since this is what many other researchers have found. Subsequently, we found that the dummy variable indicating a zero test score was significantly positively associated with higher attainments! Individuals who get zero test scores may be rebellious, possible bored with school, but they are not stupid.

[5] Other researches have found that attitudes towards school affect test scores. This result was first noted in Coleman *et al* (1966) but has been confirmed in a number of papers since then.

students who like school do better at it and produce better test scores than those who do not. But liking school is not the same thing as being smart.

It is even harder to associate the class quality variable with any measure of innate ability. Most Danish parents of high ability children could not arrange for their children to be in classes where most of the other children in the class were also high ability either by switching schools or by getting a class change for their child. The significance of this variable is not the result of a selection process in which class ability and individual ability are synonymous. Peer effects matter because fourteen year olds like to conform and if the norm is high educational achievement then individuals in the class will perform closer to their potential than would be the case if the reverse were true.

We are not the only researchers to find variables like these that play an important role in explaining test score variation. Zavodny (2005) found significant correlations between standardized test score performance and hours of television watched in most of her model specifications. Feinstein and Symons (1999: 309) obtain a large and highly significant coefficient for their peer group variable. Segal (2005: 21) finds that the relation between $8^{th}$ grade misbehaviour and test score performance is of the same order of magnitude as that between family background variables and test scores. Heckman and Rubenstein (2001: 148) find that previous involvement in illicit activities is correlated with test score performance with the direction of the effect being determined by the subgroup being considered. Finally, Lipscomb (2007) found that participation in school sponsored clubs and sports activity increased math and science test scores.

Individuals who have high levels of latent ability - individuals who are 'smart'- will do well on test scores if they are not disaffected, are disciplined and highly motivated and have acquired the skills to deal with the abstraction involved with testing procedures. In the context our test score production model these individuals have above average amounts of human capital as well as a desire to do well. Likewise, individuals with no latent ability will do poorly. But very smart individuals may do poorly because they are not interested, have behavioural problems, come from families or attend schools where the culture

places a low value on learning and ability, or for one reason or another have never learned to apply their abilities to abstract problem solving. Because of this it the smartest person in the class may not get the highest test score result.

It is clear from our results and from what others have found that intelligence tests do not measure intelligence or 'fundamental cognitive ability' but a very large number of attributes which are inherited, acquired from, or imposed on by the individual's family, school, or social environment and are, thus, a broad measure of the individual's human capital. Earlier studies like those of Korenman and Winship (2000) focused on the potential importance of socioeconomic variables that affected the individual when he or she was a child. To this list we add cultural variables, in line with Fryer and Levitt (2004), attitude and school quality variables and suggest that the accuracy of test score data may be compromised by incentive compatibility problems. As more research is done the list of variables which affect test scores continues to grow. Where it will end and exactly what role 'fundamental cognitive ability' will play in their determination remains to be seen.

# 6 Acknowledgment

The authors wish to thank John Geweke for insightful comments on an earlier version as well as the participants of the session on the economics of education at the meetings of The European Economic Association in Vienna in August 2006.

# References

[1] Albernaz, Angela; Ferreira, Francisco H. G.; Franco, Creso (2002). "Qualidade e equidade no ensino fundamental brasileiro". (With English summary.) *Pesquisa e Planejamento Economico* 32: 453-76.

[2] Arrow, Kenneth, Bowles, Samuel; and Durlauf, Steven (2000). *Meritocracy and Economic Inequality.* Princeton, N.J.: Princeton University Press

[3] Ben Porath, Y (1967). "The Production of Human Capital and the Life Cycle of Earnings" *Journal of Political Economy* 75: 352-365.

[4] Björklund, Anders, Lindahl, Mikael and Plug, Erik (2006). "The Origins of Intergenerational Associations: Lessons from Swedish Adoption Data", *Quarterly Journal of Economics* 121, 999-1028.

[5] Brody, Nathan (1992) Intelligence 2nd ed. Academic Press, San Diego, CA.

[6] Coleman, James S., Campbell, E.Q., Hodgson, C.J., Mood, J., Weinfield, F.J. and York, R.L. (1966). Equality of Educational Opportunity US. Government Printing Office, Washington D.C.

[7] Daniels, M., Devlin, Bernie and Roeder, Kathryn (1997). "Of Genes and IQ". Ch. 3 in Devlin *et al* (1997).

[8] Deb, Partha and Trivedi, Pravin K. (1997). Demand for Medical Case by the Elderly: a Finite Mixture Approach. *Journal of Applied Econometrics* 12: 313-336.

[9] DeFries, John C., Plomin, Robert and Fulker, David W. (1994). *Nature and Nurture During Middel Childhood*. Blackwell Press, Oxford UK.

[10] Devlin, Bernie, Feinberg, Stephen E., Resnick, Daniel P. and Roeder, Kathryn (1997). Intelligence, Genes, and Success: Scientists Respond to *The Bell Curve.* Copernicus Books.

[11] Eckland, Bruce K. (1979). Genetic Variance in the SES-IQ Correlation. *Sociology of Education* 52, No. 3.: 191-196.

[12] Feinstein, Leon and Symons, James (1999). Attainment in secondary school. *Oxford Economic Papers* 51: 300-321.

[13] Fryer, Roland and Levitt, Steven D. (2004). "Understanding the Black-White Test Score Gap in the First Two Years of School" *Review of Economics and Statistics* 86: 447-464.

[14] Gordon, Margaret T. (1976). A Different View of the IQ-Achievement Gap. *Sociology of Education* 49, No. 1: 4-11.

[15] Greene, William H. (2003). Econometric Analysis, Prentice Hall, fifth edition.

[16] Hansen, Erik J. (1995). *En generation blev voksen*, report 95: 8 (English summary 1996) The Danish National Institute of Social Research, Copenhagen

[17] Hansen, Karsten T., Heckman, James J., and Mullen, Kathleen J. (2004). "The Effect of Schooling and Ability on Achievement Test Scores" *Journal of Econometrics* 121: 39-98.

[18] Herrnstein, Richard and Murray, Charles (1994). *The Bell Curve*. New York: Free Press.

[19] Heckman, James J. and Singer, B. (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica* 52: 271-320

[20] Heckman, James J. and Rubenstein, Yona (2001). The Importance of Non-Cognitive Skills: Lessons from the GED Testing Program" Papers and Proceedings, *American Economic Review*. 91: 145-149.

[21] Hunt, Earl J. (1997). "The Concept of Utility of Intelligence". Ch. 7 in Devlin *et al* (1997).

[22] Jencks, C., Smith, M., Acland, M., Bane, M. Cohen, D., Gintis, H., Heyns, B. and Michelson, S. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York: Basic Books.

[23] Korenman, Sanders and Winship, Christopher (2000). A Re-analysis of The Bell Curve: Intelligence, Family Background, and Schooling, In: Arrow, KJ, Bowles, S, Durlauf, S (eds) *Meritocracy and Economic Inequality* Princeton University Press, Princeton, 137-178

[24] Lambert, D. (1992). "Zero inflated Poisson regressions, with an application to defects in manufacturing." *Technometrics* 34: 1-14.

[25] Marais, M. Laurentius and Wecker, William E. (1998). Correcting for Omitted-Variables and Measurement-Error Bias in Regression with an Application to the Effect of Lead on IQ. *Journal of the American Statistical Association* 93, iss. 442: 494-505.

[26] Lipscomb, Stephen (2007). "Secondary School Extracurricular Involvement and Academic Achievement: A fixed Effects Approach" *Economics of Education Review*, 26, 463-472.

[27] McIntosh, James and Munk, Martin D. (2007). Scholastic Ability vs. Family Background in Educational Success: Evidence form Danish Sample Survey Data, *Journal of Population Economics*, 20.

[28] Mullahey, J. (1986). "Specification and Testing Some Modified Count data Models. *Journal of Econometrics* 33: 341-365.

[29] Neal, Derek A. and Johnson, William R. (1996). "The Role of Pre-Market Factors in Black-White Wage Differences" *Journal of Political Economy* 104: 869-895.

[30] Peters, H. Elizabeth and Mullis, Natalie C. (1997). "The Role of Family Income and Sources of Income in Adolescent Achievement", Pp. 340-381 in *Consequences of Growing Up Poor*, edited by G.C. Duncan and Jean Brooks-Gunn, New Yor: Russel Sage Foundation.

[31] Plomin, Robert, Fulker, David W., Corley, Robin and DeFries, John C. (1997). "Nature, Nurture, and Cognitive Development From 1 To 16 Years". *Psychological Science* 8: 442-447.

[32] Plug, Erik and Vijverberg, Vim (2005). "Does Family Income Matter For Schooling Outcomes? Using Adoptees As a Natural Experiment." *Economic Journal* 115: 879-906.

[33] Sacerdote, Bruce (2004). "What Happens When We Randomly Assign Children To Families?" Unpublished Manuscript, Dartmouth College.

[34] Scarr, Sandra and Weinberg, Richard A. (1978), the influence of "family background" on intellectual attainment. *American Sociological Review* 43: 67-92.

[35] Segal, Carmit (2005). Misbehavior, Education and Labor Market Outcomes" Unpublished manuscript, Stanford Economics Department.

[36] Steelman, Lala Carr and Mercy, James A. (1983). Sex Differences in the Impact of the Number of Older and Younger Siblings on IQ Performance. *Social Psychology Quarterly* 46, No. 2.: 157-162.

[37] Todd, Petra E. and Wolpin, Kenneth I. (2003). "On the Specification and Estimation of the Production Function for Cognitive Achievement" *Economic Journal* 113: F3-F33.

[38] _____ (2004) "The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps" PIER Working Paper #04-019.

[39] Greene, William H. (2003), *Econometric Analysis*, Prentice Hall, fifth edition.

[40] Wedel, M., Desarbo, W.S., Bult, J.R. and Ramaswamy, V. (1993). "A Latent Class Poisson Regression Model for Heterogeneous Count Data" *Journal of Applied Econometrics* 8: 397-411.

[41] Winship, Christopher and Korenman, Sanders (1997). "Does Staying in School Make You Smarter? The Effect Education on IQ in *The Bell Curve*". Ch. 10 in Devlin *et al* (1997)

[42] Zajonc, R. B. and Markus, G. B. (1975). Birth order and intellectual development. *Psychological Review* 82:74-88.

[43] Zavodny, Madeline (2005). "Does Watching Television Rot Your Mind? Estimates of the Effect on Test Scores" *Economics of Education Review,* 25, 565-573.

# Tables

**TABLE 1**

**Summary Statistics For The Inductive Reasoning Test Scores**

| Summary Statistic | Boys | Girls |
|---|---|---|
| Mean | 21.78 | 21.85 |
| Standard Deviation | 9.10 | 9.61 |
| Minimum | 0 | 0 |
| Maximum | 39 | 40 |
| Percent Zero | 3.4 | 4.0 |
| Quartile 1 | 24.2 | 25.4 |
| Quartile 2 | 21.4 | 20.6 |
| Quartile 3 | 24.3 | 25.4 |
| Quartile 4 | 30.1 | 28.5 |
| Sample Size | 1557 | 1577 |

## TABLE 2

### Variable Means and (Standard Deviations)

| Variable | Boys | | Girls | |
|---|---|---|---|---|
| **School Variables** | | | | |
| Likes school | 0.39 | (0.47) | 0.52 | (0.50) |
| Indifferent to school | 0.52 | (0.49) | 0.30 | (0.48) |
| Dislikes school | 0.09 | (0.29) | 0.18 | (0.38) |
| School class quality | 0.36 | (0.48) | 0.33 | (0.47 |
| **Household Variables** | | | | |
| Income | 30.36 | (16.88) | 30.75 | (16.80) |
| Mother home | 0.37 | (0.48) | 0.37 | (0.48) |
| Financial problems | 0.25 | (0.43) | 0.20 | (0.40) |
| Broken home | 0.13 | (0.34) | 0.11 | (0.31) |
| Number of siblings | 2.13 | (1.50) | 2.05 | (0.03) |
| Urban | 0.24 | (0.43) | 0.25 | (0.43) |
| **Father's Occupation** | | | | |
| Skilled white and blue collar workers | 0.29 | (0.33) | 0.29 | (0.46) |
| Professional and managerial | 0.49 | (0.42) | 0.46 | (0.50) |
| Unskilled | 0.22 | (0.15) | 0.25 | (0.41) |
| **Parents' Education** | | | | |
| Father's education | 0.63 | (0.49) | 0.60 | (0.49) |
| Mother's education | 0.79 | (0.49) | 0.78 | (0.49) |

## TABLE 3

### Model Selection Criteria

| Number of Distributions | Number of Parameters | Boys | Girls |
|---|---|---|---|
| | | ln(L) | ln(L) |
| 1 | 1 | -7353.401 | -7880.467 |
| 1 | 14 | -6960.736 | -7575.201 |
| 2 | 29 | -5767.951 | -5982.312 |
| 3 | 44 | -5446.275 | -5587.507 |

**TABLE 4**

**Mixed Maximum Likelihood Parameter Estimates**

**For The Inductive Reasoning Test Score By Type For Boys.**

| | **Type I** | **Type II** | **Type III** | **Unmixed** |
|---|---|---|---|---|
| **Household variables** | | | | |
| Constant term | 3.270** (0.028) | 2.673** (0.053) | 1.404** (0.394) | 3.062** (0.010) |
| Family income | 0.022† (0.013) | -0.024 (0.027) | -1.225** (0.284) | 0.017** (0.006) |
| Mother home | 0.001 (0.017) | 0.083** (0.032) | 1.076** (0.345) | 0.019** (0.007) |
| Financial problems | 0.031† (0.019) | 0.102** (0.035) | 2.239** (0.367) | 0.046** (0.008) |
| Broken home | 0.002 (0.023) | 0.053 (0.043) | 0.980** (0.353) | 0.005 (0.009) |
| Number of siblings | -0.013** (0.004) | -0.041** (0.009) | -1.033* (0.463) | -0.047** (0.001) |
| Urban | -0.009 (0.019) | 0.026 (0.039) | -0.195 (0.345) | -0.032** (0.008) |
| **School Variables** | | | | |
| Likes School | 0.009 (0.016) | 0.080** (0.030) | 1.368** (0.308) | 0.028** (0.007) |
| Dislikes School | -0.125** (0.028) | -0.158** (0.056) | -0.130** (0.033) | -0.188** (0.011) |
| Teacher's evaluation | 0.056** (0.015) | 0.077** (0.028) | -1.033* (0.463) | 0.085** (0.006) |
| **Father's occupation** | | | | |
| Managerial | 0.044** (0.021) | 0.211** (0.040) | 0.522 (0.356) | 0.128** (0.008) |
| Skilled | -0.051* (0.024) | -0.039 (0.044) | -2.625** (0.611) | -0.040** (0.009) |
| **Parents'education** | | | | |
| Father's education | 0.012 (0.017) | 0.071* (0.033) | 0.586† (0.347) | 0.027** (0.007) |
| Mother's education | 0.053** (0.022) | 0.053** (0.041) | -0.547 (0.372) | 0.057** (0.008) |
| **Type Characteristics** | | | | |
| Probability of type | 0.606** (0.018) | 0.343** (0.018) | 0.051** (0.006) | - |
| Conditional probability | 0.000 | 0.035 | 0.964 | - |
| Predicted mean | 27.108 (2.117) | 14.946 (2.794) | 3.002 (9.819) | 21.777 (3.283) |

†, ∗, and ∗∗ indicate significant at the 10, 5, and 1 percent levels, respectively.

Standard errors are in round brackets.

**TABLE 5**

**Mixed Maximum Likelihood Parameter Estimates**

**For The Inductive Reasoning Test Score By Type For Girls.**

| | **Type I** | **Type II** | **Type III** | **Unmixed** |
|---|---|---|---|---|
| **Household variables** | | | | |
| Constant term | 3.274** (0.026) | 2.552** (0.051) | 2.864.** (0.204) | 3.075** (0.009) |
| Family income | 0.019 (0.015) | 0.033 (0.097) | -0.288* (0.121) | 0.008 (0.005) |
| Mother home | 0.013 (0.018) | 0.037 (0.032) | 0.498** (0.143) | 0.028** (0.006) |
| Financial problems | 0.027† (0.017) | 0.064† (0.039) | 0.750** (0.180) | 0.012† (0.008) |
| Broken home | -0.040† (0.024) | 0.053 (0.043) | -3.549** (0.869) | -0.036** (0.024) |
| Number of siblings | -0.008* (0.004) | -0.099* (0.045) | -0.575** (0.047) | -0.038** (0.001) |
| Urban | -0.0162 (0.019) | -0.044 (0.035) | -0.388* (0.180) | -0.007 (0.007) |
| **School Variables** | | | | |
| Likes School | 0.027 (0.017) | 0.003** (0.032) | -0.279† (0.151) | 0.005 (0.006) |
| Dislikes School | -0.088** (0.021) | -0.144** (0.041) | -1.845** (0.235) | -0.125** (0.008) |
| Teacher's evaluation | 0.064** (0.016) | 0.124** (0.031) | 1.074** (0.161) | 0.104** (0.006) |
| **Father's occupation** | | | | |
| Managerial | 0.086** (0.022) | 0.211** (0.040) | 0.171 (0.195) | 0.120** (0.008) |
| Skilled | 0.003 (0.024) | 0.067 (0.045) | -1.286** (0.252) | -0.010 (0.009 ) |
| **Parents' education** | | | | |
| Father's education | 0.022 (0.017) | 0.079* (0.034) | -0.513** (0.182) | 0.057** (0.006) |
| Mother's education | 0.019 (0.021) | 0.041 (0.040) | -1.054** (0.199) | -0.009 (0.007) |
| **Type characteristics** | | | | |
| Probability of type | 0.576** (0.018) | 0.343** (0.017) | 0.080** (0.008) | - |
| Conditional probability | 0.000 | 0.061 | 0.939 | - |
| Predicted mean score | 28.045 (2.246) | 15.220 (2.485) | 4.260 (6.852) | 21.851 (2.897) |

†, ∗, and ∗∗ indicate significant at the 10, 5, and 1 percent levels, respectively.

Standard errors are in round brackets.

**TABLE 6**

**Goodness Of Fit Statistics**

**(Actual / Predicted)**

| **Variable** | **Boys** | **Girls** |
|---|---|---|
| Mean | 21.41 / 21.37 | 21.85 / 21.51 |
| Standard Deviation | 7.95 / 7.81 | 9.10 / 8.74 |
| %=0 | 3.4 / 3.4 | 3.4 / 3.2 |
| %=5 | 0.4 / 2.2 | 0.6 / 0.5 |
| %=10 | 1.5 / 4.0 | 1.6 / 1.4 |
| %=15 | 3.3 / 3.6 | 3.5 / 3.5 |
| %=20 | 4.2 / 4.6 | 3.7 / 3.8 |
| %=25 | 5.3 / 5.3 | 4.2 / 4.3 |
| %=30 | 3.3 / 3.3 | 3.3 / 3.4 |
| %=35 | 1.3 / 1.0 | 2.3 / 1.6 |
| %=40 | 0.0 / 0.1 | 0.0 / 0.4 |
| Psuedo $R^2$ | 25.9 | 29.1 |