What Do Test Scores Really Mean? A Latent Class Analysis of Danish Test Score Performance

Munk, Martin D.; McIntosh, James

Published in:
Scandinavian Journal of Educational Research

DOI (link to publication from Publisher):
10.1080/00313831.2013.773554

Publication date:
2014

Document Version
Accepted author manuscript, peer reviewed version

Scandinavian Journal of Educational Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/csje20

What Do Test Score Really Mean? A Latent Class Analysis of Danish Test Score Performance

James McIntosh[a] & Martin D. Munk[b]
[a] Economics Department, Concordia University
[b] Department of Political Science, Aalborg University Copenhagen
Published online: 29 Apr 2013.

To cite this article: James McIntosh & Martin D. Munk (2014) What Do Test Score Really Mean? A Latent Class Analysis of Danish Test Score Performance, Scandinavian Journal of Educational Research, 58:4, 435-452, DOI: 10.1080/00313831.2013.773554

To link to this article: http://dx.doi.org/10.1080/00313831.2013.773554

Scandinavian Journal of Educational Research, 2014
Vol. 58, No. 4, 435–452, http://dx.doi.org/10.1080/00313831.2013.773554

Routledge
Taylor & Francis Group

What Do Test Score Really Mean? A Latent Class Analysis of Danish Test Score Performance

James McIntosh
Concordia University

Martin D. Munk
Aalborg University

Latent class Poisson count models are used to analyse a sample of Danish test score results from a cohort of individuals born in 1954–1955, tested in 1968, and followed until 2011. The procedure takes account of unobservable effects as well as excessive zeros in the data. 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 socioeconomic status of the parents and 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 and possible incentive problems make it more difficult to understand what the tests measure.

Keywords: human capital, educational production functions, test scores, ability, unobservable heterogeneity and types

Issues in Educational Testing

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 when the respondents are adolescents have been shown to be highly significant variables in models which explain later educational attainments as well as success in labour markets.[1]

James McIntosh, Economics Department, Concordia University; Martin D. Munk, Department of Political Science, Aalborg University Copenhagen.

The authors wish to express their thanks to Nikolai Gospodinov for helpful discussion as well as the five anonymous referees for their comments.

Correspondence concerning this article should be addressed to M. D. Munk, Centre for Mobility Research, Department of Political Science, A. C. Meyers Vaenge 15, Building: Frederikskaj 10B, Aalborg University Copenhagen, DK-2450 Copenhagen SV, Denmark. E-mail: mdm@dps.aau.dk

[1]See Heckman (2008, p. 300) for the United States, and McIntosh and Munk (2007, p. 114) for Denmark.

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.

Some educational researchers, Herrnstein and Murray (1994, p. 22) for example, refer to educational test scores taken at early ages 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."; others such as Flynn (2007, p. 2) refer to items in Raven's Progressive Matrices Test to be part of an IQ test by stating that "…a single generation on an IQ test of forty items selected from Raven's Progressive Matrices". Not everyone agrees with the Herrnstein and Murray position. 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.

Similar views are put forward by Richardson (2002) in his review of the literature on what test scores mean.

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 (see, e.g., Cicmanec, 2009). If, for example, it is mistakenly believed that test scores primarily measure innate intelligence then individuals who do poorly on these tests could be labelled 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 focussed on remedial programmes that help disadvantaged students overcome their problems.

One of the purposes of this research is first to review the literature on test score determinants to see what it has to say about how test performance is determined. Although the question that we pose is not often considered explicitly,[2] 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 analyse the results from an inductive reasoning test that was administered to a representative sample of Danish students aged fourteen in 1968. Which variables are the most important in determining test performance? How important are family background variables in determining how well respondents perform? How important are unobservable effects and what might they represent? These are some of the questions that this research will address.

The data comes from The Danish Longitudinal Survey of Youth (DLSY) 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 of correct answers to the test which we analyse using count models.

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

Our preferred specification is the Latent Class Poisson Model of Wedel, Desarbo, Bult and Ramaswamy (1993) which takes account of the presence of excessive zeros as well as dealing directly with the problem unobservable respondent characteristics. This procedure allows respondents with the same observable characteristics to respond to them differently because of differences in unobservable attributes. This approach is new to the analysis of test scores.

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 the scholastic abilities of the respondent's classmates as measured by the teacher's assessment of the respondent's class or the number of siblings that the respondent had, variables which have nothing to do with the innate ability of the individual, were also highly significant covariates in the test score model. Unobservable effects as represented by the typology of respondents that is generated by the application of mixture distributions in latent class models were more important than those associated with observable variables, in particular for the two first types. Some respondents refused to answer any of the test questions suggesting the presence of incentive problems. As a result it is more difficult to determine how well testing procedures can actually elicit what test scores actually measure. This leads to a rather different perspective on the meaning of test score results. Our results also help us to interpret the role of genetic or biological factors in the determination of test outcomes.

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 reflect a process which is evolving. 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 and provides a brief discussion of value added models. Section 3 describes the data and gives a detailed description of the statistical models used in the analysis. The results are developed in section 4 and section 5 discusses their implications and concludes.

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, 2007) examine test scores in an educational production framework first suggested by Ben-Porath (1967). Here achievements or test scores are related to the histories of two input vectors by assuming

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

In equation (1) $T_i(a_i)$ is the test score of individual i at age a, and 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, Heckman and Mullen (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. $\mu_i(0)$ and f_i are what we have been referring to as "innate" ability. In this formulation achievement or manifest ability, as measured by the test score, is determined by a zero mean IQ variable f_i mediated by a scaling factor, $\lambda(s_i)$, 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 the 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. Consequently, it is essential that procedures which deal explicitly with the presence of unobservables be employed in the statistical analysis of test score data.

Most of the research which examines test score performance relies on data which is fairly limited in scope. Todd and Wolpin (2007) found that having observable investments as well as the ages at which they were made mattered as far as the results were concerned. They showed that value added models, which use a test score administered at an earlier age as a regressor, could be used to deal with the problems that arise from missing information. Although there are limitations to this procedure it is quite widely used. Feinstein and Symons (1999) apply a value added model to the British National Development Survey which contains all children born in Britain between 3 and 9 March 1958. 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 model also provides additional insight into the true nature of test scores which suggests to us that measured ability is dynamic and evolves over time and reflects the effects of schooling and continuing parental investments as the Todd–Wolpin model suggests rather than some unchanging innate level of pure intelligence. 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 (2012a) 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.

Value added procedures require observations on a test score at two different ages. Unfortunately, the data used in this study has only one test score observation so it will not be possible to use a value added model in the analysis. However, it will be interesting to see how our results compare with studies that utilize this additional information.

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) analyse 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 30% are significant.

Other earlier studies in chronological order are Zajonc and Markus (1975), Gordon (1976), Scarr and Weinberg (1978), Eckland (1979), Paulhus and Shaffer (1981), Steelman and Mercy (1983), Neal and Johnson (1996), Peters and Mullis (1997), Albernaz, Ferreira and Franco (2002), Diseth (2002), Zwick and Himelfarb (2011), Kempe, Eriksson-Gustavsson and Samuelsson (2011), and finally Hansen, Rosén and Gustafsson (2011). A common finding in all of these studies, as well as the papers surveyed in Sackett, Kuncel, Arneson, Cooper and Waters, (2009), is the importance of socioeconomic status as represented by parent incomes, occupations or educational attainments.

Finally, some of our results relate to the nature–nurture debate. There is a very large literature on this topic (e.g. Daniels, Devlin and Roeder, 1997 for a survey). Causal results usually flow from studies involving twins or adopted children. Even here there is considerable diversity in the results that psychology, epigenetics, and behavioural genetics researchers have found. While many traits, including those involving cognitive skills, depend on both environmental and genetic characteristics, as Neisser et al. (1996) note, there is considerable disagreement on the proportion that is genetically inherited from parents. Bouchard and McGue (2003, p. 4) take the view that "…there is now strong evidence that virtually all individual psychological differences, when reliably measured, are moderately to substantially heritable", a position which is also held by Plomin, Fulker, Corley and DeFries (1997). On the other hand Petronis (2010, p. 722) interprets the empirical evidence from twin studies as more favourable to environmental factors because of the high proportion of personality trait variation due to non-shared environmental factors.

Methods

In this section, the data, summary statistics, and the statistical models are presented.

Data and Summary Statistics

The data comes from a survey originally carried out by E.J. Hansen (1995). Information on 3151 subjects was collected during the period 1968/1969. The data includes taxable family income, the employment status of the subject's mother (i.e. whether or not she remained at home during his/her childhood), marriage status of the subject's parents, the number of siblings, the teacher's evaluation of the school class to which the respondent attended when the

Table 1
Summary Statistics For The Inductive Reasoning Test Scores

Summary Statistic	Boys	Girls
Mean	21.8	21.9
Standard Deviation	9.1	9.6
Minimum	0	0
Maximum	39	4
Percent Zero	3.4	4.0
Quartile 1	11.6	12.2
Quartile 2	30.4	30.4
Quartile 3	38.9	35.9
Quartile 4	19.1	21.4
Sample Size	1557	1577

tests were taken, 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[3] which contained 40 questions designed to elicit the respondent's ability to perform logical operations. The questions are number theoretic; the respondents are asked to complete a sequence of numbers using the pattern in the sequence.[4] The students had 18 minutes to answer all the 40 questions (Ørum, 1971, p. 26). The score is the number of correct answers obtained on the test. Table 1 contains the relevant summary statistics for the third test. Among the subjects there were 17 students (0.54% of the sample) whose test scores were not observed; these subjects were not included. There are also 116 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 respond to the test questions in a way which accurately reflects their true ability to answer them.

Table 2 contains information on the variables which are used to explain the test scores. 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 average school score is also included as a regressor. 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

[3]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 analyse 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 McIntosh and Munk (2007) so it was also excluded.

[4]The test is available on the Web site: http://www.sfi.dk/dokumentation-6790.aspx.

Table 2
Variable Means and Standard Deviations

Variable	Boys		Girls	
School Variable				
Teacher's evaluation of 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)
Number of siblings	2.13	(1.50)	2.05	(0.03)
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)

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 14 to 15. 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 variable 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.

Statistical Models

The most popular way of dealing with test score data has been the use of ordinary least squares. This is consistent with classical test theory. See Schmidt and Embretson (2003), and Bollen (2002, p. 611). In this methodology the actual score, T_i, is assumed to be equal to the true score, T_i^*, plus an error term, ε_i, or

$$T_i = T_i^* + \varepsilon_i. \tag{3}$$

It has a traditional regression representation when $T_i^* = X_i\beta$.

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 score data. For our data, however, this procedure is not particularly appropriate. It does not address the problems of excessive zeros. Table 1 shows that the test has nearly a 4% zero response for both boys and girls.

The respondents with these zeros actually took the tests but did not get any correct answers. Although this percentage is 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 and the Poisson–Gamma model requires that all respondents have test scores which depend on their family background variables in exactly the same way. There are two reasons for rejecting the Poisson–Gamma model. The first, as Heckman and Singer (1984) noted, the regression parameter estimates depend on an arbitrary choice of the mixing distribution. Secondly, the distribution of the Gamma error does not depend on the covariate vector X_i. This problem can be resolved by using a latent class model which allows the errors to be correlated with the respondent's family characteristics. We think that this is an advantage since some of what is missing probably depends on family characteristics.

The procedure used here, which we now outline, focusses 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 to the latent class models developed by Wedel, Desarbo, Bult and Ramaswamy (1993), (see also Cameron & Trivedi, 2005, p. 678–679; Skrondal & Rabe-Hesketh, 2007; Winkelmann, 2008). We assume that there are a finite number of types each with a different set of unobservable characteristics. We assume that if respondent i is of type ℓ, i will respond to the covariates, X_i, which describe his or her observable characteristics in a way which depends on these unobservable type ℓ characteristics. This is our way of allowing respondents with the same observable characteristics to respond differently to them because of differences in their unobservable characteristics. The expected test score for an individual i who is of type ℓ, is then defined as

$$E(T_{i\ell}) = \exp\left[X_i\beta_\ell\right] = \mu_{i\ell}, \tag{4}$$

where β_ℓ is determined ℓ by type unobservable characteristics[5] and $X_i\beta_\ell = \beta_{\ell 0} + \sum_{K=1}^{K} \beta_{\ell k} x_{ik}$. The intercept term for type ℓ can be seen as a measure of unobserved type ℓ ability. It affects the test score directly. But type ℓ ability also determines what respondent i gets from variables like parental educational attainments, for example.

For each type, the score is assumed to have a Poisson distribution and the probability of being type ℓ is p_ℓ. The likelihood function for this type of model can be found in the references listed above. 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 an intercept term.[6]

[5]McIntosh and Munk (2007, p. 110) show that writing $E(T_{i\ell}) = \exp[X_i\beta + \epsilon_\ell]$, where ϵ_ℓ is an error which depends on X_i and takes a finite number of values, is equivalent to the model described by equation (4).

[6]Our estimates are derived by maximizing the likelihood function. However, practitioners who want to use packaged programmes can use Latent Gold 4.5 which is written by Vermunt and Magidson

Table 3
Model Selection Criteria

Number of Distributions	Number of Parameters	In(L)
1	1	−15,223.963
1	19	−14,576.432
2	36	−11,792.938
3	59	−11,061.176
3	46	−11,068.967

This serves as a baseline which can be used to compare other models and to construct a pseudo-R^2.[7] Additional mixtures were added until there was no significant increase in the penalized likelihood function or until convergence difficulties were encountered.

Much of the psychological and educational testing literature focusses on the analysis of the individual items in the test rather than the score which is being used here which is the sum of the correct answers. The procedures used in the analysis of the individual items come under the rubric of item response theory for which a good reference is van der Linden and Hambleton (1997).[8]

Results

Parameter estimates appear in Tables 4 and 5 by gender and type. Because there are a number of respondents at each school we use White's (1980) robust standard errors to deal with cluster effects.[9] These are used instead of the usual error to deal with any clustering effects that might arise because there is more than one respondent from each school.

It is clear from Table 3 that the likelihood function continues to increase significantly as the number of mixtures increases. 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. But three mixtures are clearly superior to the unmixed model in terms of AIC or BIC criteria or likelihood ratio tests. This does not mean that there are only three distinct types; there could be more but the data is not rich enough and the sample sizes are too small to be able to estimate models with a larger number of distinct types.

Gender is important and was dealt with by using the following procedure. First, the two genders were estimated separately. This is an inefficient way of estimating the parameters.

(2008). Poisson models also belong to the generalized linear model family and there is a considerable amount of software written for these models.

[7]The pseudo-R^2 is the percentage increase in the ln-likelihood function over baseline. It is also known as McFadden's R^2.

[8]Latent class analysis is sometimes applied to models which explain the results of the individual items. Lindsay, Clogg and Grego (1991) is an example.

[9]Hierarchical linear models, like those described in Raudenbush and Bryk (2002), are sometimes used to deal with clustering. However our models are not linear and these models do not deal with unobservables. These models are also equivalent to random effects models, see Greene (2008, p. 223), so that using a robust estimator will correct for the heteroscedasticity induced by clustering.

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								
Intercept term	3.278**	(0.034)	2.494**	(0.070)	−2.377*	(1.064)	3.030**	(0.024)
Family income	0.005	(0.005)	0.005	(0.005)	0.007	(0.227)	0.006	(0.05)
Mother home	−0.005	(0.016)	−0.005	(0.016)	0.581*	(0.291)	0.014	(0.012)
Number of siblings	−0.010**	(0.003)	−0.037**	(0.010)	−0.455*	(0.181)	0.042**	(0.003)
School Variables								
Teacher's evaluation of school class quality	0.059**	(0.012)	0.059**	(0.012)	0.099	(0.366)	0.087**	(0.011)
School average test score	0.003**	(0.001)	0.007**	(0.002)	0.280	(0.284)	0.005**	(0.001)
Father's Occupation								
Managerial and Professional	0.066**	(0.018)	0.333**	(0.046)	1.412*	(0.693)	0.138**	(0.015)
Skilled white and blue collar workers	−0.021	(0.024)	0.090†	(0.052)	0.270	(0.829)	−0.019	(0.017)
Parents' Education								
Father's education	0.047*	(0.023)	0.047*	(0.023)	0.974†	(0.606)	0.070**	(0.023)
Mother's education	0.054**	(0.022)	0.022	(0.043)	1.026	(1.286)	0.058*	(0.025)
Type Characteristics								
Probability of type	0.593**	(0.015)	0.347**	(0.014)	0.061**	(0.005)	1.0	
Predicted mean	27.245**	(1.999)	15.440**	(2.760)	3.083**	(3.710)	21.668**	(2.161)

†,*, and ** indicate significant at the 10%, 5%, and 1% levels, respectively. Robust standard errors are in round brackets.

However it did reveal that both the intercept terms and type probabilities were the same for both genders. These restrictions were imposed by pooling the data but allowing separate parameters for both gender and type. The value of the maximized likelihood function for this model is −11,061.176, as indicated in the fourth row of Table 3. However, 13 of the 59 coefficients were not significantly different across gender or type and were, therefore, constrained to be equal. These constraints satisfy a likelihood ratio test. For the restricted model there are no obvious gender differences for types I and II. However, imposing the same coefficients for both genders for types I and II was rejected by a likelihood ratio test. For type III boys have significantly different coefficients for household income, father having a professional occupation, mother's education, and the number of siblings.

In terms of goodness of fit statistics the model fits the data well. The pseudo-R^2 is 0.273. This is quite high for sample survey data. The zeros and the first two moments of each distribution are well predicted and, as Figure 1 demonstrates, the model's predictions accurately track the cumulative scores. Using mixtures allows all of the model to fit the data very well.

These models also accurately predict the number of zeros. Latent class mixture models were first used by Deb and Trivedi (1997) to deal with the problem of excessive zeros. Deleting the zeros generates the selection problem noted by Heckman (1979) which will lead to parameter biases. 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.

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								
Intercept term	3.278 **	(0.034)	2.494**	(0.070)	−2.378**	(1.064)	3.030**	(0.024)
Family income	0.005	(0.005)	0.005	(0.005)	1.017**	(0.291)	−0.003	(0.020)
Mother home	0.007	(0.017)	0.007	(0.017)	0.407	(0.393)	0.023†	(0.012)
Number of siblings	−0.010**	(0.003)	−0.018	(0.011)	−0.747**	(0.181)	−0.036**	(0.00)
School Variables								
Teacher's evaluation of school class quality	0.059**	(0.011)	0.059**	(0.011)	−0.357	(0.594)	0.098**	(0.011)
School average test score	0.003**	(0.001)	0.007**	(0.012)	0.047**	(0.017)	0.005**	(0.001)
Father's Occupation								
Managerial and Professional	0.074**	(0.020)	0.261**	(0.050)	−0.864	(0.602)	0.119**	(0.015)
Skilled white and blue collar workers	0.000	(0.024)	0.158**	(0.054)	−0.087	(0.436)	−0.002	(0.016)
Parents' Education								
Father's education	−0.045**	(0.017)	−0.045**	(0.017)	1.924**	(0.912)	0.084**	(0.012)
Mother's education	−0.035†	(0.020)	−0.101*	(0.054)	−2.032**	(0.473)	−0.006	(0.014)
Type Characteristics								
Probability of type	0.593**	(0.015)	0.347**	(0.014)	0.061**	(0.005)	1.0	
Predicted mean score	27.934**	(1.964)	14.892**	(2.130)	2.089	(5.410)	21.805**	(2.049)

†,*, and ** indicate significant at the 10%, 5%, and 1% levels, respectively. Robust standard errors are in round brackets.

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 and they have the highest predicted mean. While their scores depend on whether their class at school was a good one, the number of siblings and the father having a high level job, and at least one of the parent education variables, the coefficients are much smaller than the coefficients for these variables for the other two types. It is the intercept 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 the respondents in this group appear to be severely disaffected. Type III boys and girls respond negatively to their parents' income and the number of siblings they have. They do not benefit being in a school class where performance levels are high but benefit from a mother who does not work. For these respondents the intercept term is very small and it is probable that the non-responders or zero scores are of this type. Type II respondents are an intermediate case. They are more sensitive to their family backgrounds but have a lower intercept than Type I respondents' term indicating a lesser importance of unobservable effects on their performance.

The importance of this result is that it shows that individuals with akin parental types can respond differently to their environments. This is not surprising. Parents with the same

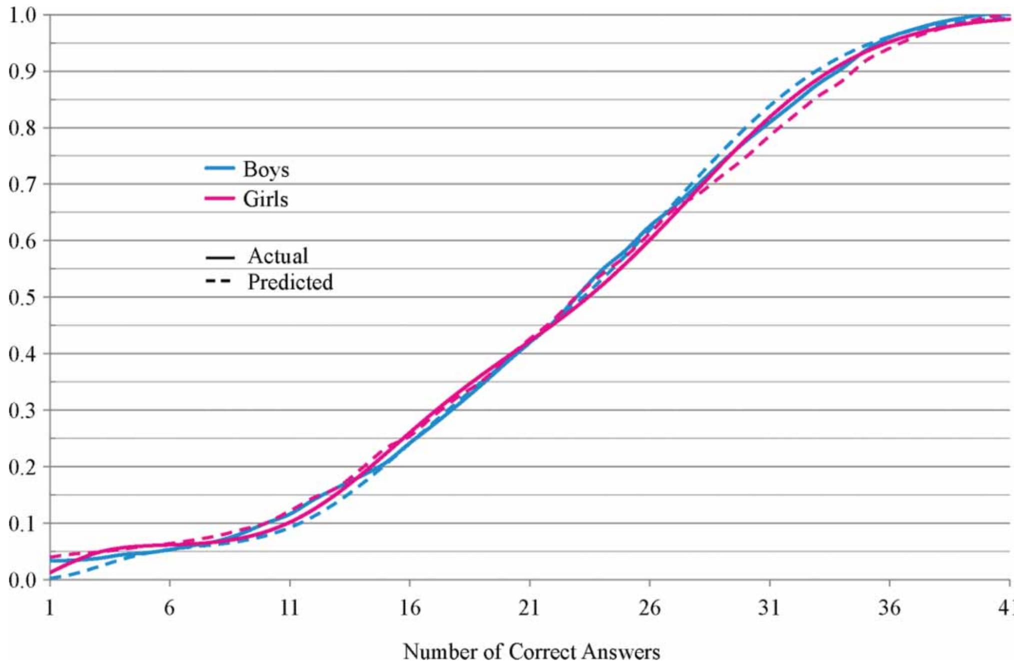

Figure 1. Actual vs. Predicted Cumulative Scores

education or occupation can be very different and they can provide different types of advantage or disadvantage for their children. Children also have different attitudes and personalities. Affluent well-educated 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 and we have not encountered anything like it in the educational test score or behavioural genetics literature. This is caused by the relatively large value of the intercept term for Type I respondents.

Intercept terms pick up the mean effects of variables which matter but can not be observed by the researcher.[10] How is this to be interpreted? One possibility is that the intercept terms are picking up the effects of early parental investments in the child that Todd and Wolpin mention or some other benefits that the respondents get from the household in which they resided as children and adolescents.[11] These variables could reflect what parents actually have done for their children rather than who the parents are in terms of their educational

[10]In a regression model it is possible to write the equation as $y_i = \alpha_\ell X_{i\ell} + \beta U_{i\ell}$ for a type ℓ respondent. In this equation $X_{i\ell}$ and $U_{i\ell}$ are the observed and unobserved variables associated with respondent i, respectively. Add and subtract $\beta_\ell \bar{U}_\ell$ to get $y_i = \beta_\ell \bar{U}_\ell + \alpha_\ell X_{i\ell} + \beta_\ell \left(U_{i\ell} - \bar{U}_\ell\right)$. It is clear that the intercept term, $\beta_\ell \bar{U}_\ell$, picks up the mean effects of variables that can not be observed. There are also unobservable effects due to the error term and the fact that this relationship is specific to each type. Of course, a regression model is not the model being used here but the principle applies.

[11]The importance of lagged test scores in the value added papers referred to in the literature review support this interpretation.

or occupational characteristics, but may not have a very high correlation with the parent's occupations or educational attainments. They could also represent characteristics which are external to the family such as those related to the respondent's school, neighbourhood, or peers. In any case what the intercept terms represent are independent of any household specific effect: they are imposed similarly to all households.

Earlier we mentioned that there are some variables which are significant in explaining test scores but are not associated with the respondent's innate ability. The class quality variable is one of them. 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 14 year olds are likely 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.[12] The number of siblings that the respondent has is another example.

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, p. 309) obtain a large and highly significant coefficient for their peer group variable as we do. Segal (2012a) finds that the relation between 8th grade misbehavior and test score performance is of the same order of magnitude as that between family background variables and test scores. Heckman and Rubenstein (2001, p. 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. Lipscomb (2007) found that participation in school sponsored clubs and sports activity increased math and science test scores. Finally, Rønning (2011) notes that the amount of homework assigned in school affects Norwegian test scores.

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 score since it is almost impossible to get all of the answers on all three tests wrong even if respondents had randomly selected the answers to the questions. 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 a 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.[13]

[12]Ammermueller and Pischke (2009) find a moderate effect of peer socioeconomic background on reading test scores in six European countries (a somewhat stronger effect than found in previous studies).

[13]Incentives to perform well are also related to the motivation of the respondent. Other researchers have noted the importance of this. A recent paper by Segal (2012b) shows that motivation seems to play an important role in test score performance.

Discussion and Conclusion

The result on the dominance of the intercept terms is an important one because it suggests that the role that biology or genetics plays in what children get from their parents is more difficult to determine than some researchers who have contributed to the "nature–nurture" debate have suggested. These intercept terms can not represent a household specific genetic transfer from parent to child since they represent the mean effects of omitted variables and have to be the same for all respondents. Researchers who examine surveys involving adopted children, Björklund, Lindahl and Plug (2006), Plug and Vijverberg (2005), and Sacerdote (2004), find a larger role for biological than environmental factors.[14] Results obtained by behavioral geneticists like Plomin et al. (1997, p. 444)[15] leads 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 behavioral 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 for adoptees is, of course, absurd.

The question at issue here is how much does the innate ability of respondent i ($\mu_i(0)$ or f_i in the notation of Todd and Wolpin (2003) and Hansen et al. (2004)) depend on the genetic characteristics of i's parents? If children inherit significant amounts of their ability from their parents one way which this could happen is for higher levels of parental ability to be reflected in better parental academic performance or more prestigious occupations. The strength of the mechanism by which the attributes which made the parents successful are passed onto their children is measured by the significance of these parental attainment and occupational variables as regressors. These serve as proxies for inherited ability in the test score model. The problem here is that parental educational attainments and occupations play a minimal role in explaining test scores.[16] Hence our results suggest that the claim that genetic factors are more important than environmental factors in determining educational success needs to be treated with some caution.

---

[14]Björklund, Eriksson and Jäntti (2010) using data on twins present results which are more in line with what we are reporting. They used the sibling correlation rather than the parent–child correlation as an estimate of the strength intergenerational associations. Whereas they find that only about 13% of the variation in son's test score can be accounted for by father's test score, the share is about 50% when using the sibling test score. This suggests that parental aspirations, attitudes and parenting practices may be important in accounting for the child's performance, leaving open the role for parent genetic characteristics.

[15]See also DeFries, Plomin and Fulker (1994).

[16]The pseudo-R^2 is around 0.04 when only family background variables are used as regressors. This increases dramatically to 0.27 when latent class models are used.

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 of our test score production model these individuals have some innate ability, 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 behavioral 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 the smartest person in the class may not get the highest test score result.

We found that there is a hierarchy of respondents. Latent class models allow us to identify different types of individuals who may be subject to different external factors and who respond to their personal circumstances and family backgrounds in different ways. The dominance of the positive intercept terms for Type I and II and a negative intercept term for Type III respondents suggests that there are important factors which lie beyond the respondent’s household or family background which determine test performance. These results are new and are consequences of our application of latent class procedures to the analysis of test score data. It is clear from our results and from what others have found that what are commonly referred to as IQ tests do not measure just intelligence or “fundamental cognitive ability” but a very large number of attributes which lead to productive outcomes. These attributes, skills, or qualities can be inherited or acquired from the respondent’s parents, learned at school from teachers and peers, or gained by participating in various social activities. As the value added studies referred to earlier indicate, they can evolve as the respondent gets older. Earlier studies like those of Korenman and Winship (2000) focussed on the potential importance of socioeconomic variables that affected the individual when he or she was a child and adolescent (see also Heckman 2006). To this list we add cultural variables, in line with Fryer and Levitt (2004), parent attitude and school quality variables. 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 (see, e.g., Munk, 2013).

References

Albernaz, A., Ferreira, F.H.G., & Franco, C. (2002). Qualidade e equidade no ensino fundamental Brasileiro. (With English summary.). *Pesquisa e Planejamento Economico*, *32,* 453–476.

Ammermueller, A., & Pischke, J.-S. (2009). Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics*, *27*(3), 315–348.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy*, *75,* 352–365.

Björklund, A., Lindahl, M., & & Plug, E. (2006). The origins of intergenerational associations: Lessons from Swedish adoption data. *Quarterly Journal of Economics*, *121,* 999–1028.

Björklund, A., Eriksson, K.H., & Jäntti, M. (2010). IQ and family background: Are associations strong or weak? *The B.E. Journal of Economic Analysis & Policy*, *10*(1), art. 2.

Bollen, K.A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53,* 605–634.

Bouchard, T.J., & McGue, M. Jr. (2003). Genetic and environmental influences on human psychological differences. *Journal of Neurobiology*, *54,* 4–45.

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego, CA: Academic Press.
Cameron, C., & Trivedi, P.S. (2005). *Microeconomics, methods and applications*. Cambridge: Cambridge University Press.
Cicmanec, K.M. (2009). Review of Measuring up: What educational testing really tells us. *Journal of Educational Measurement*, *46,* 363–366.
Daniels, M., Devlin, B., & Roeder, K. (1997). Of genes and IQ. In B. Devlin, S.E. Feinberg, D.P. Resnick, & K. Roeder (Eds.), *Intelligence, genes, and success: Scientists respond to the bell curve*, (Ch. 3). New York: Copernicus Books.
Deb, P., & Trivedi, P.K. (1997). Demand for medical case by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, *12,* 313–336.
DeFries, J.C., Plomin, R., & Fulker, D.W. (1994). *Nature and nurture during middle childhood*. Oxford: Blackwell Press.
Diseth, Å. (2002). The relationship between intelligence, approaches to learning and academic achievement. *Scandinavian Journal of Educational Research*, *46,* 221–230.
Eckland, B.K. (1979). Genetic variance in the SES-IQ correlation. *Sociology of Education*, *52,* 191–196.
Feinstein, L., & Symons, J. (1999). Attainment in secondary school. *Oxford Economic Papers*, *51,* 300–321.
Flynn, J.R. (2007). *What is intelligence? Beyond the Flynn effect*. New York: Cambridge University Press.
Fryer, R., & Levitt, S.D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, *86,* 447–464.
Gordon, M.T. (1976). A different view of the IQ-achievement gap. *Sociology of Education*, *49,* 4–11.
Greene, W.H. (2008). *Econometric analysis* (6th ed.). New York: Prentice Hall.
Hansen, E.J. (1995). *En generation blev voksen*, Report 95: 8 (English summary 1996). Copenhagen: The Danish National Institute of Social Research.
Hansen, T., Heckman, J.J., & Mullen, K.J. (2004). The effect of schooling and ability on achievement test scores. *Journal of Econometrics*, *121,* 39–98.
Hansen, K.Y., Rosén, M., & Gustafsson, J.-E. (2011). Changes in the multi-level effects of socio-economic status on reading achievement in Sweden in 1991 and 2001. *Scandinavian Journal of Educational Research*, *55,* 197–211.
Herrnstein, R., & Murray, C. (1994). *The bell curve*. New York: Free Press.
Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, *47,* 153–161.
Heckman, J.J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900–1902.
Heckman, J.J. (2008). Schools, skills, and synapses. *Economic Enquiry*, *46,* 289–324.
Heckman, J.J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, *52,* 271–320.
Heckman, J.J., & Rubenstein, Y. (2001). The importance of non-cognitive skills: Lessons from the GED testing program. Papers and proceedings. *American Economic Review*, *91,* 145–149.
Kempe, C., Eriksson-Gustavsson, A.-L., & Samuelsson, S. (2011). Are there any Matthew effects in literacy and cognitive development? *Scandinavian Journal of Educational Research*, *55,* 181–196.
Korenman, S., & Winship, C. (2000). A re-analysis of the bell curve: Intelligence, family background, and schooling. In K.J. Arrow, S. Bowles, & S. Durlauf (Eds.), *Meritocracy and economic inequality* (pp. 137–178). Princeton, NJ: Princeton University Press.
Lambert, D. (1992). Zero inflated Poisson regressions, with an application to defects in manufacturing. *Technometrics*, *34,* 1–14.
Lindsay, B., Clogg, C.C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*(413), 96–107.

Lipscomb, S. (2007). Secondary school extracurricular involvement and academic achievement: A fixed effects approach. *Economics of Education Review*, *26,* 463–472.

McIntosh, J., & Munk, M.D. (2007). Scholastic ability vs. family background in educational success: Evidence from Danish sample survey data. *Journal of Population Economics*, *20*(1), 101–120.

Mullahey, J. (1986). Specification and testing some modified count data models. *Journal of Econometrics*, *33,* 341–365.

Munk, M.D. (2013). Completion of upper secondary education: What mechanisms are at stake? *Comparative Social Research*, *30,* 255–291.

Neal, D.A., & Johnson, W.R. (1996). The role of pre-market factors in black-white wage differences. *Journal of Political Economy*, *104,* 869–895.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51,* 77–101.

Paulhus, D., & Shaffer, D.R. (1981). Sex differences in the impact of number of older and number of younger siblings on scholastic aptitude. *Social Psychology Quarterly, 44*, 363–368.

Peters, H.E., & Mullis, N.C. (1997). The role of family income and sources of income in adolescent achievement. In G.C. Duncan & J. Brooks-Gunn (Eds.), *Consequences of growing up poor* (pp. 340–381). New York: Russel Sage Foundation.

Petronis, A. (2010). Epigenitics as a unifying principle in the aetiology of complex traits and diseases. *Nature*, *465,* 721–727.

Plomin, R., Fulker, D.W., Corley, R., & DeFries, J.C. (1997). Nature, nurture, and cognitive development from 1 to 16 years. *Psychological Science*, *8,* 442–447.

Plug, E., & Vijverberg, V. (2005). Does family income matter for schooling outcomes? Using adoptees as a natural experiment. *Economic Journal*, *115,* 879–906.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks: Sage Publications.

Richardson, K. (2002). What IQ tests test. *Theory & Psychology*, *12*(3), 283–314.

Rønning, M. (2011). Who benefits from homework assignments? *Economics of Education Review*, *30,* 55–64.

Sacerdote, B. (2004). *What Happens When We Randomly Assign Children To Families?* Unpublished Manuscript. Dartmouth College.

Sackett, P.R., Kuncel, N.R., Arneson, J.J., Cooper, S.R., & Waters, S.D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin*, *135,* 1–22.

Scarr, S., & Weinberg, R.A. (1978). The influence of "family background" on intellectual attainment. *American Sociological Review*, *43,* 67–92.

Schmidt, K.M., & Embretson, S.E. (2003). Item response theor. In J.A. Schinka, W.F. Velicer, & I.B. Weiner (Eds.), *Handbook of psychology: Research methods in psychology* (Ch. 17). New York: Wiley.

Segal, C. (2012a), Misbehavior, Education and Labor Market Outcomes. *Journal of the European Economic Association*. E-PUB ahead of print.

Segal, C. (2012b). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, *58*(8), 1438–1457.

Steelman, L.C., & Mercy, J.A. (1983). Sex differences in the impact of the number of older and younger siblings on IQ performance. *Social Psychology Quarterly*, *46,* 157–162.

Skrondal, A., & Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, *34,* 712–745.

Todd, P.E., & Wolpin, K.I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, *113,* F3–F33.

Todd, P.E., & Wolpin, K.I. (2007). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human Capital*, *1,* 91–136.

van der Linden, W.J., & Hambleton, R.K (Eds.). (1997). Item response theory: Brief history, common models, and extentions. *Handbook of modern item response theory* (Ch. 1). New York: Springer.

Vermunt, J.K., & Magidson, J. (2008). *LG-Syntax user's guide: Manual for Latent GOLD 4.5 Syntax Module*. Belmont, MA: Statistical Innovations Inc.

Wedel, M., Desarbo, W.S., Bult, J.R., & Ramaswamy, V. (1993). A latent class Poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, *8,* 397–411.

White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, *48,* 817–838.

Winkelmann, R. (2008). *Econometric analysis of count data*. Berlin: Springer-Verlag.

Zajonc, R.B., & Markus, G.B. (1975). Birth order and intellectual development. *Psychological Review*, *82,* 74–88.

Zavodny, M. (2005). Does watching television rot your mind? Estimates of the effect on test scores. *Economics of Education Review*, *25,* 565–573.

Zwick, R., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement*, *48,* 101–121.

Ørum, B. (1971). *Social baggrund, intellektuelt niveau og placering i skolesystemet*, Study 20. Copenhagen: The Danish National Institute of Social Research.