Using Cognitive Test Scores in Social Science Research

Lynne Steuerle Schofield

Thesis Proposal

June 10, 2007

Abstract

A standard problem in social science attempts to better understand the large wage disparities between black and white workers in U. S. labor markets. Social scientists have conducted hundreds of studies of observed racial wage gaps, seeking to understand the extent to which they are driven by differences in human capital or disparate treatment by employers. In order to get an unbiased estimate of such effects, it is necessary to include in the regression equations measures of human capital. While years of schooling has traditionally been used as a measure of human capital, social scientists are increasingly turning to cognitive test scores, as a more direct measure. Most social science research that uses cognitive test scores as an independent variable models the test score as fixed and without error. However, since test scores have measurement error, modeling the test score in this way can produce biased results which can result in incorrect policy conclusions. Current methods for modeling the test score with error are limited to single point in time analysis with a fixed cognitive assessment administered to all subjects, and situations in which the measurement error is homogeneous across all subjects. In response to these drawbacks, a new model called the Mixed Effects Structural Equations (MESE) model is developed. The MESE model is demonstrated using data from the National Adult Literacy Survey by analyzing black-white wage gaps in married men, single men, and single women. Three important findings are of note. First, much of the black-white wage gap can be attributed to a black-white disparity in skills suggesting that more attention ought to be focused on the development of skills. Second, comparisons of the the MESE model to a model with no measurement error demonstrate the importance of modeling the measurement error. Third, comparisons of the MESE model to a model using current methodology suggest the MESE model may solve some of the drawbacks noted in the other current methods.

1 Introduction

In their landmark paper, Neal and Johnson (1996) use test scores from the Armed Forces Qualifying Test (AFQT) as a measure of human capital[1] to better understand whether continued wage differentials among minorities and whites are a result of labor force discrimination or differences in skills. Their paper is well cited (Google Scholar boasts more than 300 citations) as it was one of the first to argue that years of school as a measure of human capital was "less than satisfactory" (p. 870) and suggested instead for the use of test scores to serve as the proxy variable. The literature documents numerous other tests and assessments that have been used by social scientists as proxies for human capital, skills, or intelligence. Besides the AFQT, the General Educational Development (GED) credential (Tyler, Murnane, and Willett, 2000), the National Longitudinal Study of the High School Class of 1972 (NLS72) and the High School and Beyond (HSB) assessment (Murnane, Willett, and Levy, 1995) and the National Adult Literacy Survey (Raudenbush and Kasim, 1998 and Venezky and

[1] I use Deardorff's Glossary of International Economics (2000) definition of human capital as "the knowledge and skill, that results from education, training, and experience, that makes an individual more productive."

Kaplan, 1998) have been used in analyses that are regression-based. In most of these papers, the ultimate goal is to provide data-based guidance for public policy. Given the implications that may be drawn from such work, it makes sense to carefully consider the way in which the cognitive test score is used as an independent variable.

In most cases involving test scores, the research places the test score as the *dependent* variable in the analysis. However, there is a modest and growing literature in the social sciences in which the test score serves as an *independent* variable.

While the literature is replete with guides for social scientists about how to estimate test scores and their standard errors, there is little research to guide the social scientist in how to use the test score as an independent variable in an analysis. Most researchers simply analyze the test score as a fixed variable. This can be problematic since the psychometric literature makes it clear that test scores are imperfect measures. All test scores have error. Thus, an analysis that includes a cognitive test score as a covariate without modeling the error will lead to biased estimates of the coefficients. In particular, when the test score is correlated with other covariates in the analysis (like race in Neal and Johnson's work), the estimates on the regression coefficients for those other covariates will be biased as well. This can lead to incorrect policy conclusions.

My work has four major goals. First, I will critically review the literature from item response theory, errors-in-variables analysis, missing data analysis, and latent variable analysis. Second, based on the review, I will develop the Mixed Effects Structural Equations (MESE) model to analyze cognitive test scores as independent variables in regressions and appropriately account for the measurement error that is associated with the test scores. The MESE model will overcome some of the statistical limitations of the techniques used to date. The properties of this new model will be examined, and recommendations for how and when to use the model will be given. Third, I will demonstrate the model with real-world data sets by examining the issue of black-white wage gaps in single women, single men, and married men. This demonstration will contribute to the literature on labor markets, particularly the literature on labor markets for women.

This proposal document continues as follows. Section 2 describes Item Response Theory as a way to estimate a test score and its error from responses to a test. Section 3 examines three current methods to solving the measurement error problem associated with using cognitive test scores as an independent variable: the classic errors-in-variables solution, a solution using multiple imputation called Plausible Values Methodology, and a solution using marginal maximum likelihood procedures. The latter two practices are advocated by the National Center for Education Statistics for use with some of its surveys. Section 4 develops the Mixed Effects Structural Equation (MESE) model that attempts to solve some of the problems with the methods discussed in Section 3. Section 5 provides preliminary results of applying the MESE model to a set of real-world data, and Section 6 discusses future avenues of research for the dissertation.

2 Item Response Theory

In the 1950s, most educational measurement models were based on classical test theory, which allowed psychometricians to estimate a person's ability or true score based on her performance on a test. While this theory advanced the field of educational testing greatly, scores generated from classical test theory were dependent on the test given. A person's ability was defined entirely by the test that one took. A difficult test often gave examinees low ability estimates whereas if the test was easy, the examinees appeared to have a higher ability. There was a desire to have a group of measurement models that were not test dependent (Hambleton, Swaminathan, and Rogers, 1991).

Psychometricians advanced a new measurement system called item response theory (Lord, 1980) to address the many issues with classical test theory. Under item response theory (IRT), the typical approach to measuring ability consists of constructing a test with a number of questions or items in which each item measures some aspect of the ability of interest. The relationship between an examinee's ability and an examinee's responses on a test is based on an item response function (IRF) also called the item characteristic curve (ICC). The IRF is a monotonically increasing function which shows that as an examinee's ability increases so does the likelihood of answering an item correctly (Baker, 2001).

A standard IRT model is the three parameter logistic (3-PL) model (Hambleton, Swaminathan, and Rogers, 1991). In this model the probability that individual j responds correctly to item x_i, given a cognitive ability θ_j, a "discrimination parameter" a_i, a "difficulty parameter" b_i, and a "guessing parameter" c_i is

$$P(x_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_j - b_i)}} \tag{1}$$

where x_{ij} is the correctness of person j's response to item i. with equal to 1 when the item is correctly answered and 0 when the item is incorrectly answered.

When individual skill estimates are needed, maximum likelihood procedures are often used, where the IRT likelihood is

$$L(x_1, x_2, \ldots, x_m|\theta_1, \theta_2, \ldots, \theta_n) = \prod_j \prod_i P(x_i|\theta_j)^{x_i}(1 - P(x_i|\theta_j))^{1-x_i}. \tag{2}$$

Methods for estimating the IRT model include the marginal maximum likelihood practice of first estimating the item parameters using E-M methods and then taking the item parameters to be known and fixed at their calibrated values when proceeding with inference regarding θ (Hambleton, Swaminathan, and Rogers, 1991) and more recently Markov Chain Monte Carlo (MCMC) methods which estimate item and subject parameters at the same time (Patz, Junker, 1999).

As noted in the introduction and underscored in the previous section, test scores are imperfect measures. Any analysis using them will need to take this measurement error into account. In the next section, I discuss three current methods for modeling the test score as an independent variable while including the measurement error, and discuss the shortcomings of these methods.

3 Current Practices in Using Cognitive Test Scores as an Independent Variable

3.1 Classic Errors-in-Variables Regression Analysis

The classic errors-in-variables regression literature discusses how to model error in the independent variables in a regression (Anderson, 1984). This literature demonstrates that when a covariate has measurement error, but is modeled without error, that the estimated coefficients will be inconsistent. The coefficient on the variable that is measured with error will be biased toward 0. In addition, when the variable measured with error is correlated with other covariates in the analysis, the coefficients on those variables will also be biased (Cheng and Van Ness, 1999).

In his study of black-white wage gaps, Bollinger (2003) shows that analyses that employ cognitive test scores as measured without variability can lead to incorrect policy conclusions. He compares the results of two models. The first model is an ordinary least squares (OLS) model in which cognitive skills (as measured by the AFQT) are modeled as fixed effects.

$$\begin{aligned} w &= \beta_0 + \beta_1\theta + \beta_2 B + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2) \end{aligned} \tag{3}$$

In the second model, he uses a classic errors-in-variables regression analysis where the AFQT score is modeled as a proxy variable that is linearly related to human capital,

$$w = \beta_0^* + \beta_1^*\phi + \beta_2^* B + \varepsilon \tag{4}$$

$$\begin{aligned} \phi &= \delta + \Gamma\theta + \nu \\ E(\epsilon|\phi, B) &= 0 \\ Cov(\nu, \epsilon) &= 0 \end{aligned} \tag{5}$$

where w is log wages, θ is the true test score, ϕ is the estimated test score, B is an indicator variable that is 1 when the person is Black and 0 when the person is White. Bollinger found that when cognitive skills are modeled without error that the estimates of the coefficients on both θ and B are biased.

While Bollinger aptly demonstrates the bias of OLS estimates when measurement error is not modeled, his model is problematic. Bollinger's model, like most classic errors-in-variables models, is unidentifiable. The errors-in-variables literature documents six assumptions that will make the model identifiable: knowing σ_ε^2, knowing σ_ν^2, knowing the ratio $\sigma_\varepsilon^2/\sigma_\nu^2$, knowing both σ_ε^2 and σ_ν^2, knowing β_0, or knowing the reliability ratio $\sigma_\varepsilon^2/(\sigma_\nu^2 + \sigma_\varepsilon^2)$ (Cheng and Van Ness, 1999).

In order to estimate his model, Bollinger places bounds on the slope parameter. (Bollinger follows Klepper and Leamer (1984) to estimate his bounds. The lower bound is the coefficient when using a direct regression or when the measurement error is assumed to be zero. The

upper bound is the reciprocal of the regression coefficient from a reverse regression or when the regression error is assumed to be zero.) With these bounds, Bollinger is able to estimate ranges for the coefficients on race and AFQT score. However, without further information or assumptions, point estimates of the coefficients are not possible with the classic errors-in-variables model.

The errors-in-variables model is problematic in this case for an additional reason. Model 2, by using the classical errors-in-variables assumptions uses true score theory to model the error of the test scores. True score theory postulates that the observed score is a linear function of the true score and random error (Lord and Novick, 1968). In true score theory, the error is assumed to be normal. This implies that the magnitude of the standard error for each individual will be the same in Model 2. However, when comparing the scores of individuals using IRT, the magnitude of the standard error can vary depending on the number and variability of the items given as well as how closely the difficulty of the items match the ability of the individual. A model that uses IRT instead of true score theory to model the test scores will allow the size of the measurement error to be different for each individual, which is likely to be desired, since individual test scores often have very different measurement error depending on the number and variability of the items on a test.

3.2 Statistics Depending on Missing Data

Two other current approaches to modeling the error inherent in the test scores use IRT instead of true score theory. Both use a multi-level Bayesian model with the IRT likelihood included as one of the levels. In both cases, the Bayesian model assumes that the test score is missing for all people in the survey. The following section introduces the theory behind the multi-level Bayesian model used in the Plausible Values Methodology and the marginal maximum likelihood model.

Consider a survey where v is the collection of survey variables. Researchers want to calculate some statistic $s(V)$ (i.e., the mean of the variable V.) When v is missing for some portion of the sample, Rubin (1977) suggests partitioning v into (v_{mis}, v_{obs}) in order to calculate $s(v)$ through its conditional expectation

$$E[s(V)|V_{obs}] = \int s(V_{mis}, V_{obs}) p(V_{mis}|V_{obs}) dV_{mis} \tag{6}$$

(Note that other statistics, like the conditional mode, the conditional median, regression coefficients, etc. are also options.)

In the Plausible Values methodology and the model using marginal maximum likelihood procedures, the survey variables can be grouped into three categories: the true test score for each individual, θ, items on a cognitive test, x, and questions that concern demographic or background information about the individual, y. We will consider the test score missing for each individual. We are interested in the statistic $s(\theta, X, Y)$ where X and Y are known. Specifying (6) to our particular case, the conditional expectation of $s(\theta, X, Y)$ is

$$s(X, Y) \quad = \quad E[s(\theta, X, Y)|X, Y]$$

$$= \int s(\theta, X, Y) p(\theta|X, Y, \alpha, \gamma) d\theta, \tag{7}$$

where, in what follows, α = parameters in the 'hyperprior' for θ, and γ = parameters in the likelihood for $X|\theta$.

Using Bayes Theorem

$$p(\theta|X, Y, \alpha, \gamma) = \kappa_{\alpha,\gamma} p(X|\theta, Y, \alpha, \gamma) p(\theta|Y, \alpha, \gamma). \tag{8}$$

Separating (8) into two pieces allows us to simplify it. Using two standard assumptions in IRT: 1) responses to items on a test are only dependent on the latent ability and 2) individual responses are independent of one another, given the latent ability $p(X|\theta, Y, \alpha, \gamma)$ can be written

$$\begin{aligned} p(X|\theta, Y, \alpha, \gamma) &= p(X|\theta, \gamma) \\ &= \prod_{j=1}^{M} \prod_{i=1}^{N} p(X_{ij}|\theta_j, \gamma). \end{aligned} \tag{9}$$

Additionally, $p(\theta|Y, \alpha, \gamma)$ can be simplified by assuming that θ is not dependent on γ and each individual's ability is independent of another's,

$$\begin{aligned} p(\theta|Y, \alpha, \gamma) &= p(\theta|Y, \alpha) \\ &= \prod_{j=1}^{M} p(\theta_j|Y_j, \alpha) \end{aligned} \tag{10}$$

leaving the following measurement model that is a multilevel/Bayes model of the form

$$\begin{aligned} \theta &\sim p(\theta|Y, \alpha) \\ X &\sim p(X|\theta, \gamma) \end{aligned} \tag{11}$$

where the IRT model is encoded in the likelihood $p(X|\theta, \gamma)$.

This measurement model is used in both the Plausible Values Methodology and the methodology that uses marginal maximum likelihood procedures. There are two main differences in how these two methodologies approach estimation for this model 1) what covariates Y_j to include in the conditioning model (10) and 2) how to calculate the integral (7).

3.2.1 The Conditioning Model

Mislevy (1991) examined the issue of what covariates, Y, must be included in (10). Several examples in his paper demonstrate that problems can occur both when Y contains too few variables and when Y contains too many variables. For reasons that will become clear below, we are interested in the case of calculating (10) when the Y variables in the posterior density for θ are different from those in $s(\theta, X, Y)$.

If the posterior for θ depends on a different set of covariates Y^* than the covariates Y in $s(\theta, X, Y)$ then (10) becomes

$$s(X, Y, Y^*) = \int s(\theta, X, Y) p(\theta | X, Y^*, \alpha, \gamma) d\theta. \quad (12)$$

$s(X, Y, Y^*)$ can be thought of as an unbiased estimator of $s(X, Y)$ when $Y \subseteq Y^*$, because then $E[s(X, Y, Y^*)|X, Y] = s(X, Y)$. Thus, it makes sense to require at least $Y \subseteq Y^*$. Mislevy (1991) gives several examples showing that $s(X, Y, Y^*)$ can exhibit arbitrary biases if $Y \not\subseteq Y^*$

In practice, we do not know exactly what Y should be and different Y's may be relevant for different $s(X, Y)$. In order to ensure that $Y \subseteq Y^*$, one might consider making Y^* very large. However, when Y contains many variables, estimates of s will be computationally inefficient.

3.2.2 Calculating the Integral (7)

Evaluating the integral in (7) will often be difficult numerically. Two methods may be used to calculate (7): multiple imputation and numerical quadrature.

Multiple imputation is often considered a way in which to deal with missing data, but it can also be considered a way to approximate an integral. In multiple imputation, random draws from $p(\theta|X, Y^*, \alpha, \gamma)$ fill the missing latent variables to give a data set with no missing values so that s can then be evaluated by averaging $s(\theta, X, Y)$ over these draws. Because approximations of the standard errors of s will be underestimated if only a single imputation is used, Rubin (1977) suggests using multiple draws. It is then possible to evaluate s repeatedly. The variance of the repeated evaluations of s will account for the variability from imputing values for the missing responses.

In numerical quadrature, the integral is approximated by summation over a set of quadrature points that fall within the range within which all observations are likely to be. When the dimension of θ is large or $s(\theta, X, Y)p(\theta|X, Y^*, \alpha, \gamma)$ is not smooth, estimates using numerical quadrature are more problematic and likely to have larger standard errors.

3.3 Current Approaches to the Measurement Error Problem using NCES surveys

3.3.1 Plausible Values Methodology: Multiple Imputation and a Saturated Conditioning Model

In the Plausible Values methodology (PV; Kirsch, I., et al., 2000), the integral is calculated using multiple imputation and the conditioning model is saturated to include as many covariates as possible. In practice, the PV methodology involves the primary analyst drawing a value from the predictive distribution $p(\theta|x_i, y_j, \hat{\alpha}, \hat{\gamma})$ (where $\hat{\alpha}$ and $\hat{\gamma}$ are Empirical Bayes estimates) for each indvidual. This step is repeated five times in order to get five "plausible values" for θ. (Note, any number of plausible values is possible.) NCES publishes the five plausible values within the dataset so that the secondary analyst can calculate estimates of

s based on the average of the five estimates

$$S_Z = \frac{1}{Z}\sum_{z=1}^{Z} s^{(Z)} \equiv \frac{1}{Z}\sum_{z=1}^{Z} s(\theta^{(z)}, X, Y) \tag{13}$$

where $Z = 5$ in the NCES examples. As in the discussion of (12) above, if $Y \subseteq Y^*$ and the θ^z are drawn from $p(\theta, X, Y, Y^*)$, then S_Z will be an unbiased of $S(X, Y)$. Thus one set of plausible values may be used by any analyst whose Y is contained by the Y^* used by the agency who generated the plausible values.

As noted above the covariates included in the conditioning model must contain the covariates used in $s(X, Y)$. Since NCES cannot know what covariates a secondary analyst may want in $s(X, Y)$, they saturate the conditioning model by making it as large as possible. Then, when the secondary analyst uses the published plausible values, S_Z will likely be unbiased. NCES and other agencies routinely produce plausible values for secondary analysts based on large Y^* for this reason.

An efficient estimate of the variance of S_Z can be calculated (assuming the $\theta^{(z)}$ are iid) as

$$V_Z = U_Z + \left(1 + \frac{1}{Z}\right) D_Z \tag{14}$$

where

$$\begin{aligned} U_Z &= \frac{1}{Z}\sum_{z=1}^{Z} u^{(z)} \equiv \frac{1}{Z}\sum u(\theta^{(z)}, X, Y); \text{ and} \\ D_Z &= \frac{1}{Z-1}\sum_{z=1}^{Z} (s^{(z)} - S_Z)^2 \end{aligned}$$

and where $u(\theta^{(z)}, X, Y)$ is a measure of the variance of the statistic s^z.

The estimate (14) incorporates both model-based uncertainty (in U_Z) and Monte-Carlo uncertainty (in D_Z). NCES uses a jackknife procedure to account for survey sampling uncertainty. Thus, the PV methodology solves the measurement error issue up the point of Monte Carlo error. (In order to decrease the amount of Monte Carlo error, one need only produce more plausible values.)

While the PV methodology works well to solve the measurement error problem, in practice it can be difficult to implement because thye typical social scientists does not have the statistical training to understand the statistics behind the PV machinery. As a result, most do not estimate s and U using the recommended procedures. Rather, they either use the first plausible value or the median of the five plausible values as a fixed test score in their analysis. (For example, see Blau and Kahn, 2005; and Green and Riddell, 2003.)

In addition, the PV methodology only works when there is a primary analyst who can produce the plausible values. While the National Adult Literacy Survey (NALS), National Asessment of Educational Progress (NAEP), and Trends in International Math and Science Study (TIMSS) have data sets which include plausible values, many cognitive tests that social scientist are using (like the AFQT, the GED, etc) do not use plausible values. Rather,

these data sets produce one estimate of the test score (generally the MLE) for each individual. Given that very few social scientists have the statistical training to understand the PV machinery well enough to estimate s and U using the recommended procedures, there are even fewer who will be able to produce plausible values themselves, making the PV methodology impossible to use in large number of cases.

3.3.2 The Marginal Maximum Likelihood Model: Numerical Quadrature and a Small Conditioning Model

In the case of a smaller conditioning model, $p(\theta_j|y_j, \alpha)$ is based on what a secondary analyst is interested in for a particular analysis. For example, if a secondary analyst wants the mean of the proficiency for blacks, $p(\theta_j|y_j, \alpha)$ becomes $p(\theta_j|race_j, \alpha)$. Numerical quadrature can be used instead of multiple imputation to calculate the integral in (7) with marginal maximum likelihood procedures.

This methodology's main advantage is that in many cases it will be statistically more efficient than the PV methodology, because the conditioning model for θ only includes what the secondary analyst wants to include. It can be problematic, however, because (7) must be recalculated for every analysis using the appropriate conditioning model based on what the secondary analyst chooses. This can be daunting for most social scientists. Currently, the AM software developed by Cohen (2002) handles this problem by calculating the integral for every analysis.

While this method is computationally more efficient, Dresher (2006) has shown that small biases occur when estimating s and using a small conditioning model. These biases come from an assumption that the distribution of θ is normal. However, in large subgroups (e.g., all blacks or all men), Dresher has shown that the distribution of θ is not actually normally distributed. When Y only includes a small number of covariates, the assumption of normality for $p(\theta|Y, \alpha)$ is much stronger than when Y includes many covariates and hence the bias becomes more substantial.

In addition, as of this writing, the MML model has only been implemented in the AM software which is not equipped to handle models in which θ is an independent variable. So, while the model underlying the AM software handles the measurement error inherent in the test score, it cannot model the test score as an independent variable. Thus, for most analysts, the MML model (as used by the AM software) can only solve problems where the test score is the dependent variable.

4 The Mixed Effects Structural Equation Model

I have developed the Mixed Effects Structural Equation (MESE) Model to address the problems described in the errors-in-variables, the Plausible Values, and the marginal maximum likelihood methodologies above. The MESE models the test score as an independent variable in the regression, models the measurement error of the test score for each individual, produces point estimates for each of the regression coefficients, and can be used for any set

of assessments or tests rather than only those tests that include a set of plausible values.

Recall that the equation of primary interest is

$$w_j = \beta_0 + \beta_1 \theta_j + \beta_2 B_j + \varepsilon_j, \tag{15}$$

The estimates of skills (θ_i) from (19) are themselves noisy, but they can be treated as random variables in a mixed-effects regression.

The basic MESE model can be written in hierarchical form as

$$\begin{aligned} \theta|Y &\sim p(\theta|Y,\alpha) && (16)\\ X|Y,\theta &\sim IRT(X|Y,\theta,\gamma) && (17)\\ w|Y,\theta &\sim N(\beta_0 + \beta_Y Y + \beta_L \theta, \sigma^2) && (18) \end{aligned}$$

where $IRT(X|Y,\theta,\gamma)$ is the IRT model used for scaling the test.

Together (16), (17), and (18) combine to give the following likelihood for the MESE model

$$L(w_1, w_2 \ldots, w_n|\theta_1,\theta_2,\ldots,\theta_m, Y_1, Y_2, \ldots, Y_m, x_{1,1}, x_{1,2}, \ldots, x_{n,m}) = \prod_j P(\theta_j|Y) \times$$
$$\prod_j \prod_i P(x_{i,j}|\theta_j)^{x_{i,j}} (1 - P(x_{ij}|\theta_j))^{1-x_{ij}} \times \prod_j \Phi(\frac{w_j - (\beta_0 + \beta_Y Y_j + \beta_L \theta_j)}{\sigma}).$$

In practice, either priors or point-estimates are additionally needed for α and γ. In addition, priors are needed for the β's. In most cases, these will be assumed to be normally distributed with mean 0. The MESE model allows the researcher to decide whether to follow the Plausible Values methodology or the MML methodology in how saturated (16) should be. However, in order to be computationally efficient, the conditioning model in the MESE model will often include only those variables in which the researcher is most interested. In most cases, (16) will be assumed to have a normal distribution. If desired, the distribution of θ for different subpopulations within the data set can have different means and variances.

Skills are modeled as a random variable, rather than a non-stochastic explanatory variable as is typical in OLS regression models. Markov Chain Monte Carlo (MCMC) machinery is used to numerically calculate the joint posterior distribution (WinBUGS software (Spiegelhalter, Thomas, and Best, 1999) is suggested as one way to implement the model in practice).

Most social scientists will be interested in the marginal posterior estimates of the βs . After integrating out the latent variable θ, these estimates can be expressed as

$$\hat{\beta} = s(X,Y) = E[s(\theta, X, Y)|X, Y]$$

where $s(\theta, X, Y)$ is the corresponding posterior estimate of β conditional on θ. Researchers can obviously choose any marginal posterior estimate they believe is appropriate.

The MESE approach addresses a number of the problems noted above. First, in the errors-in-variables solution, the magnitude of the error is the same for each individual. The

MESE model exploits the information available in the IRT model in order to allow the measurement error to be different for each individual. Because the data set includes the item responses to a number of items for each examinee, there are effectively multiple measures of the proficiency for each individual. With these multiple measures, the MESE model utilizes the IRT model to estimate both the proficiency of each individual as well as the stochastic variation inherent in the skills measurement of each examinee. In the MESE model, unlike most classic errors-in-variable solutions, the prior on θ is motivated by the data-generation process.

Second, recall that the PV methodology requires that a primary analyst produce the five plausible values for use by the secondary analyst. The PV methodology could not be used when plausible values were not produced. Many tests used in the literature (i.e., the AFQT, the GED, and the HSB assessment etc.) do not have plausible values. The MESE model, however, can be used for any test regardless of whether or not plausible values are produced, making it useful in a more general setting.

Third, the solution using marginal maximum likelihood procedures only models the test score as a dependent variable. The MESE model is an extension of the MML model because it adds an additional level to the multi-level Bayesian model in order analyze the test score as an independent variable.

5 Preliminary Results

5.1 The Economic Problem of Black-White Wage Gaps

In order to demonstrate the MESE model in practice, I have conducted an analysis examining black-white wage gaps for men and women.

In U.S. labor markets there are large wage disparities between black and white workers. Altonji and Blank (1999) using data from the March 1996 Current Population Survey (CPS) showed that in 1995 among all workers, white males earned an average hourly wage of $18.96 in comparison to black males, who earned an average hourly wage of $12.41. White females earned an average hourly wage of $12.25 in comparison to black females, who earned an average hourly wage of $10.19. Economists seek to understand the extent to which these wage gaps are driven by differences in human capital or disparate treatment by employers.

5.2 The Data: The National Adult Literacy Survey

My data come from the 1992 National Adult Literacy Survey (NALS) which includes an individually-administered household survey of 24,944 adults ages 16 and over. The NALS is comprised of two sets of questions: standard demographic questions (i.e., race, gender, labor force behavior, marital status, education, etc.) and items that measure functional literacy in three domains: prose, document, and quantitative.

Table 1 provides some demographic characteristics of the NALS sample. A few features of the data are worth noting. First, on average, white men earn more than black men and

white women earn more than black women; the black-white wage ratio among men is 0.67 and the black-white wage ratio among women is 0.90. Second, on average black adults have less education than white adults in the sample. Third, although black women and black men have similar educational attainment, black men have lower average literacy skills. White women and white men have similar literacy skills.

The NALS contains 165 items to test the literacy skills of the examinees. Because of the length of the test, it was deemed impractical to administer every item to every respondent. Instead, respondents were randomly administered a booklet which contained a representative sub-sample of approximately one-third of the full set of items. This approach makes the calculation of precise individual proficiency scores difficult. Consequently, the NALS data set does not contain individual proficiency estimates, but instead contains five plausible values per content area and individual.

5.3 Empirical Analysis of Black-White Wage Differences

In Table 2 I present estimates for four specifications. In each specification the dependent variable is log wage. (In the example for this proposal, I only include full-time workers for whom wage data are available.)

In the first specification, the explanatory variables are an indicator variable for race—equal to 1 if the respondent is black and 0 if the respondent is white—and "potential experience," which is calculated as the individual's age minus years of schooling minus six (the age at which most children enter first grade). Previous studies have indicated that the effect of work experience on earnings is non-linear; so potential experience is entered as a quartic. In the second specification, I also include the median plausible value, as is often seen in the literature by secondary analysts who do not understand the plausible value methodology. In the third specification, I use the suggested NCES procedures for finding estimates and standard errors (the mean of the estimates across the five plausible values and jackknifing for the standard errors.) The fourth specification uses the MESE model I developed to estimate the regression parameters.

There are many reasons to expect that labor market outcomes will be different for married and unmarried individuals. Most obviously, couples often specialize in their economic activity, with one individual (usually the woman) not participating in the labor market. Thus, I will follow Neal (2004a and 2004b) by estimating the regressions separately for never married adults (referred to as "single") and currently married adults. I excluded married women entirely from my analysis. In the case of men and never-married women, potential experience is a reasonable approximation of real experience. It has been shown that potential experience is not a reasonable estimate of real experience for married women, because there are large differences in work patterns between married black and married white women (Zalokar 1990, Neal 2004a, and Black, Haviland, Sanders, and Taylor 2005).

The results in Table 2 are quite amazing. The black-white log wage gap is large when I control only for experience: -0.34 for married men, -0.18 for single men, and -0.20 for single women. However, when I condition on literacy measures the estimated black-white log wage gap becomes non-significant for all three populations, suggesting that it is reasonable

to infer that much, though quite likely not all, of the race wage gap stems from the skills gap.

The analyses using the three different measures of literacy produce interesting results. When comparing the coefficient for literacy, it is necessary to equate the results due to the fact that the distribution of literacy is different for each specification. The table includes two lines which compare the changes in log wages from one standard deviation change in the literacy distribution.

As expected, when the results of the median plausible value and the NCES recommended procedure are compared, the point estimates are quite different. By using the median plausible value, I have essentially shrunk the distribution of literacy for each subpopulation which causes an the estimates of the returns to literacy to be smaller than they actually are. In addition, in the median plausible value specification, the measurement error of the test is not modeled, causing the estimates of the returns to literacy to be smaller still.

Comparing the results of the NCES recommended procedure with my MESE model, we see similar results. In each case, race is not significantly different from zero indicating no racial wage gap when I control for skills. In a sample of young men (both married and single), Neal and Johnson find that in a regression that conditions on a measure of cognitive skills (the AFQT), the estimated black-white gap is -0.07. I estimate a gap of -0.08 for married men and 0.03 for the younger single men. As Altonji and Blank (1999) note, there is relatively little work done on racial differences in wages among women. I know of no results that are directly comparable to mine. However, Black, Haviland, Sanders, and Taylor (2005), who look at the black-white gap among college-educated single childless women, controlling for fine detail in college degree and major, find a log wage gap of -0.02—quite close to my estimate of -0.007.

In addition, estimates of the coefficient on literacy indicate that the labor market returns to literacy skills are substantial. An increase of θ equal to one standard deviation of the literacy distribution results in an approximately 20 to 27 percent increase in observed wages. On the basis of the MESE estimates, the estimated market returns to literacy are similar for men and women and blacks and whites.

Model fit for each of the specifications is examined using AIC. As expected, the AIC is much larger for the MESE model than for the three other specifications. This is due to the fact that the MESE model is estimating many more parameters than any of the other three models. Further examination of model fit will follow in the dissertation.

6 Proposed Work

The example of the MESE model in use in Section 4 presents some preliminary results, but questions remain. Namely, what conditions are necessary in order for the MESE model to produce reasonable results? In addition, how do results from the MESE model work on other sets of data? As such, I propose the following agenda for completion of the dissertation.

- Propose guidelines on the number and variation in the properties of items on a test for reasonable estimates using the MESE model.

In the MESE model, I assume that I have multiple measures of cognitive ability because I have item responses to numerous items on a test. It is likely that a test with one item will not produce good results within the MESE model. Yet it is also unrealistic to expect researchers to have tests with answers to many items. Through the use of simulation and possibly some analytic work, I will produce some recommendations on the number and variation in the properties (difficulty and discrimination parameters) of the items that a given test needs in order to produce reasonable estimates for the MESE model.

- Perform a sensitivity analysis of the model assumptions

  Numerous assumptions are made in the MESE model (i.e., a normal prior for θ). I will perform a sensitivity analysis of the model assumptions in order to determine how and when the model breaks down. I will explore the sensitivity of mis-specification of the IRT model and mis-specification of the primary equation of interest. Recommendations about how sensitive the model is to different assumptions will be outlined.

- Examine model fit and compare the fit of the MESE model to the fit of other current methods.

  Section 4 presents some preliminary results on a real world data set, but it does not examine the fit of the model. I will examine the fit of the MESE model. In addition, I will compare the fit of the MESE model to the fit of other current methods in order to evaluate how much better the MESE model is. Traditional model selection criteria like AIC and BIC will be used. I will examine the residuals (likely by exploring the use of multi-level model residuals). Finally, I will explore the possibility of determining a diagnostic to evaluate the MESE.

- Examine the identifiability of the MESE model.

  As noted in Section 2.1, the classic error-in-variables model is unidentifiable unless certain assumptions are made. I will more fully examine the identifiability of the MESE model. In particular, I will explore the MESE's ability to estimate the measurement error separate from the model error. I will develop guidelines to explain the limits and operating characteristics of the MESE in solving the identifiability problem inherent in classic errors-in-variables regression.

- Provide recommendations on what posterior estimate is most appropriate in various situations.

  In Section 4, I reported the marginal posterior means of the regression estimates. The Plausible Value methodology, though, uses something much closer to the marginal posterior mode. And there may be situations in which the marginal posterior median should be used. In many cases, particularly when N is large and normality can be assumed, these estimates should be similar. However, it is possible, even when N is fairly large, that there will be selection bias in examining social issues for certain populations. For example, in section 4, I restricted the sample in my analysis to only those who work full time. However, for men, those who work full time are likely to

have much higher human capital than those who are out of the work force. Thus, the assumption of normality for their literacy scores may be inappropriate and might affect the shape of the marginal posterior. Thus, there may be cases where one posterior estimate is more appropriate than the other. I will develop recommendations on which posterior estimate is most appropriate in various situations.

- Apply the MESE model to another data set.

  In order to demonstrate the usefulness of the MESE model in other contexts and settings, I will also explore the issue of black-white wage gaps using data from the National Longitudinal Study of Youth (NLSY). This data set includes a set of numerous demographic variables as well as item responses to the Armed Services Vocational Aptitude Battery (ASVAB) test. Some comparisons of the results from the NLSY data will be made to the results of the NALS data. These comparisons will allow me also to speculate on whether there are differences in results depending upon the age at which cognitive ability is measured. Without a complete meta-analysis, of course, it will not be possible to draw firm conclusions on this matter, but it may indicate some important directions for future research as well.

References

Altonji, J. and Blank, R. (1999), "Gender and Race in the Labor Market," in *Handbook of Labor Economics, Volume 3,* eds. O. Ashenfelter and D. Card, New York: Elsevier Science Press, 3143-3259.

Anderson, T. W. (1984), "Estimating Linear Statistical Relationships," *The Annals of Statistics,* 12(1), 1-45.

Baker, F. (2001), *The Basics of Item Response Theory,* University of Maryland, College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Black, D., Haviland, A., Sanders, S., and Taylor, L. (2006), "Why Do Minority Men Earn Less? A Study of Wage Differentials among the Highly Educated," *Review of Economics and Statistics,* 88(2), 300-313.

Black, D., Haviland, A., Sanders, S., and Taylor, L. (2005), "Gender Wage Disparities among the Highly Educated," draft, Carnegie Mellon University.

Blau, F. and Kahn, L. M. (2005), "Do Cognitive Test Scores Explain Higher US Wage Inequality?" *Review of Economics and Statistics,* 87(1), 184-193.

Bollinger, C. (2003), "Measurement Error in Human Capital and the Black-White Wage Gap," *Review of Economics and Statistics,* 85(3), 578-585.

Brown, C. and Corcoran, M. (1997), "Sex-Based Differences in School Content and the Male/Female Wage Gap," *Journal of Labor Economics,* 15, 431-465.

Card, D. and Krueger, A., (1992), "School Quality and Black-White Relative Earnings: A Direct Assessment," *Quarterly Journal of Economics,* 107, 151-200.

Cheng C-L. and Van Ness J. W., (1999), *Statistical Regression with Measurement Error,* London: Arnold Publishers.

Cohen, J. (2002), *User guide: AM Statistical Software,* Washington, DC: American Institutes for Research.

Deardorff, A. V., (2000), *Deardorff's Glossary of International Economics,* http://www-personal.umich.edu/ alandear/glossary/.

Dresher, A. (2006), "Results from NAEP Marginal Estimation Research," Paper presented at the Annual Meeting of the National Council on Measurement in Education in San Francisco, CA.

Green, D. A. and Riddell, W. C. (2003), "Literacy and Earnings: An Investigation of the Interaction of Cognitive and Unobserved Skills in Earnings Generation," *Labour Economics,* 10, 165-184.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991), *Fundamentals of Item Response Theory,* Newbury Park, CA: Sage Publications.

Johnson, W. and Neal, D. (1998), "Basic Skills and the Black-White Earnings Gap," in *The Black-White Test Score Gap,* eds., Jencks, C. and Phillips, M., Washington, DC: Brookings, 480-497.

Kirsch, I., Jungeblut, A., Jenkins, L., and Kolstad, A. (1993), *Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey,* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kirsch, I., et al. (2000), *Technical Report and Data File User's Manual For the 1992 National Adult Literacy Survey,* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Klepper, S. and Leamer, E. (1984), "Consistent Sets of Estimates for Regressions with Errors in All Variables" *Econometrica,* 52, 163-183.

Lord, F. M. (1980), *Applications of Item Response Theory to Practical Testing Problems,* Hillsdale, New Jersey: Erlbaum.

Lord, F. M. and Novick, M. R. (1968), *Statistical Theories of Mental Test Scores,* Reading, MA: Addison-Wesley.

Mislevy, R. J. (1991), "Randomization-Based Inference about Latent Variables from Complex Samples," *Psychometrika,* 56, 177-196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992), "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses," *Journal of Educational Measurement,* 29(2) NAEP, 133-161.

Murnane, R. J. Willett, J. B., and Levy, F. (1995), "The Growing Importance of Cognitive Skills in Wage Determination," *Review of Economics and Statistics,* 77, 251-266.

Neal, D., and Johnson, W. (1996), "The Role of Pre-Market Factors in Black-White Wage Differences," *Journal of Political Economy,* 104, 869-895.

Neal, D. (2004a), "The Measured Black-White Wage Gap among Women is Too Small," *Journal of Political Economy,* 112, S1-S28.

Neal, D. (2004b), "The Relation Between Marriage Market Prospects and Never-Married Motherhood," *Journal of Human Resources,* 39, 938-957.

Patz, R. J. and Junker, B. (1999), "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models" *Journal of Educational and Behavioral Statistics,* 24(2), 146-178.

Raudenbush, S. W., and Kasim, R. (1998), "Cognitive Skill and Economic Inequality: Findings from the National Adult Literacy Survey," *Harvard Educational Review,* 68, 33-79.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Spiegelhalter, D. J., Thomas, A., and Best, N. G., (1999), *WinBUGS Version 1.2 User Manual,* MRC Biostatistics Unit.

Tyler, J. H., Murnane, R. J., and Willett, J. B. (2000), "Cognitive Skills Matter in the Labor Market, Even for School Dropouts," Cambridge, MA: National Center for the Study of Adult Learning and Literacy Report 15.

Venezky, R. L. and Kaplan, D. (1998), "Literacy Habits and Political Participation," In *Literacy for the 21st Century,* ed. Smith, M. C., Westport, CN: Greenwood Publishing Group.

Zalokar, N. (1990), *The Economic Status of Black Women: An Exploratory Investigation,* U.S. Commission on Civil Rights Clearinghouse Report. Washington DC: U. S. Government Printing Office.

Table 1: Sample Characteristics of the NALS

	Black Men	Black Women	White Men	White Women
N	1665	2807	7449	9404
Avg. Age	39.4	39.4	40.5	42.4
Marital Status				
Proportion never Married	0.39	0.38	0.28	0.19
Education				
Proportion Still in HS	0.06	0.04	0.04	0.03
Proportion < HS	0.29	0.29	0.13	0.14
Proportion HS	0.29	0.29	0.27	0.30
Proportion < College	0.25	0.29	0.30	0.33
Proportion College +	0.11	0.09	0.26	0.21
Literacy Skills				
Median Prose *Plausible Value*	-0.55	-0.39	0.57	0.56
Earnings of Full-Time Workers				
Avg. Weekly Wage	452.3	397.5	674.6	440.9

Table 2: Preliminary Regressions

	Married Men				Single Men				Single Women			
	Baseline	Median PV	NCES Jackknife	MESE	Baseline	Median PV	NCES Jackknife	MESE	Baseline PV	Median Jackknife	NCES	MESE
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
Race	-0.343	-0.054	-0.088	-0.082	-0.184	0.099	0.062	0.030	-0.201	0.131	0.097	-0.007
	(0.047)	(0.046)	(0.046)	(0.039)	(0.062)	(0.067)	(0.070)	(0.061)	(0.062)	(0.060)	(0.062)	(0.058)
Lit. Skills:												
Median PV		0.267				0.274				0.346		
		(0.013)				(0.026)				(0.026)		
NCES Jackknife			0.235				0.243				0.307	
			(0.015)				(0.028)				(0.031)	
MESE				0.183				0.211				0.204
				(0.011)				(0.022)				(0.024)
Experience:												
Exp	-0.093	0.0004	-0.009	0.032	-0.091	0.023	0.010	-0.051	-0.396	-0.094	-0.145	-0.089
	(0.068)	(0.063)	(0.067)	(0.061)	(0.105)	(0.098)	(0.145)	(0.103)	(0.109)	(0.099)	(0.110)	(0.104)
Exp^2	0.008	0.002	0.003	-0.001	0.008	0.0008	0.002	0.007	0.038	0.011	0.016	0.012
	(0.005)	(0.005)	(0.005)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)
Exp^3	-0.0002	-0.0001	0.0001	0.00004	-0.0002	-0.0001	-0.0001	-0.0002	-0.001	-0.0004	-0.0006	-0.0005
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0004)	(0.0003)	(0.0004)	(0.0004)	(0.0004)	(0.0003)	(0.0003)	(0.0003)
Exp^4	$2x10^{-6}$	$-8x10^{-7}$	$1x10^{-6}$	$-7x10^{-7}$	$2x10^{-6}$	$1x10^{-6}$	$1x10^{-6}$	$3x10^{-6}$	$2x10^{-6}$	$5x10^{-6}$	$7x10^{-6}$	$6x10^{-6}$
	$(2x10^{-6})$	$(2x10^{-6})$	$(2x10^{-6})$	$(2x10^{-6})$	$(5x10^{-6})$	$(5x10^{-6})$	$(5x10^{-6})$	$(5x10^{-6})$	$(4x10^{-6})$	$(4x10^{-6})$	$(4x10^{-6})$	$(4x10^{-6})$
$\mu_{\theta,Black}$		-0.241	-0.237	0.126		-0.235	-0.238	0.075		0.012	-0.021	0.284
				(0.079)				(0.109)				(0.083)
$\mu_{\theta,White}$		0.847	0.849	1.59		0.779	0.775	1.47		1.040	1.032	1.633
				(0.026)				(0.054)				(0.058)
$\tau_{\theta,Black}$		0.330	1.042	1.158		0.341	1.053	1.243		0.336	0.889	1.054
				(0.066)				(0.083)				(0.062)
$\tau_{\theta,White}$		0.360	0.909	1.118		0.363	0.912	1.123		0.361	0.893	1.094
				(0.022)				(0.042)				(0.050)
Change in 1 st dev of lit.:												
Black		0.088	0.245	0.212		0.093	0.255	0.262		0.116	0.273	0.215
White		0.096	0.214	0.205		0.099	0.221	0.237		0.125	0.274	0.223
AIC	5495	5093	5142	78080	1680	1574	1586	22950	1491	1336	1356	20360
N	2557				710				640			

Notes: Specification (a) is a baseline regression that includes only experience entered as a quartic, and an indicator variable for race. Specification (b) adds as a measure of human capital the median prose PV. Specification (c) replaces the median PV with the NCES recommended jackknifing procedure for estimating the regression. Specification (d) uses no plausible values and instead uses the MESE model to estimate the regression coefficients. Included in the sample are individuals aged 25-55, who work full-time and reported wages and who answered at least one literacy item.