CARNEGIE MELLON UNIVERSITY

MODELING MEASUREMENT ERROR WHEN USING COGNITIVE TEST SCORES IN SOCIAL SCIENCE RESEARCH

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

In

STATISTICS AND PUBLIC POLICY

by

LYNNE STEUERLE SCHOFIELD

Department of Statistics
and
Heinz School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

July, 2008
Abstract

In many areas of social science, researchers want to use latent measures of ability as independent variables. Often cognitive test scores are used to measure this latent trait. Many social scientists do not model the measurement error inherent in the test score. I introduce the Mixed Effects Structural Equations (MESE) model to model the measurement error when a cognitive test score is used as a measure of ability as an independent variable. Unlike the typical linear regression model, which ignores the error and produces biased regression coefficients, the MESE model assumes measurement error. Unlike the typical errors-in-variables (EIV; Anderson, 1984) model which uses classic test theory (CTT) to model homoskedastic measurement error by ability, the MESE model uses item response theory to model heteroskedastic measurement error by ability. The IRT model handles the well-known identifiability issues of the EIV model. While the Plausible Value Methodology which is a marginal estimation procedure (Mislevy, 1991) produces consistent regression coefficients, a primary analyst is required to produce a set of “plausible values” for use by the secondary analyst. Inconsistent estimates can occur using plausible values if data used in the regression equation is collected after the plausible values are produced. The MESE model can be used with any test or assessment and any data set. A number of simulation studies explore the sensitivity of the MESE model assumptions, noting in particular that the prior on ability must be conditioned on the covariates in the regression equation in order to avoid bias in the estimate of the regression coefficients. The MESE model is also used to examine the issue of black-white wage gaps. I show that estimates using the MESE model differ markedly from estimates when an elementary linear regression is used. I find that most, though not all, of the black-white wage gap is plausibly attributed to the black-white
skills gap. I also explore the hypothesis that education and ability must be included in the model and find some evidence to support this claim. I find no evidence to support the hypothesis that the return to skills is unequal across race.
Acknowledgements

There are many people who deserve thanks for the help and support they have given me throughout my graduate career. First, I would like to thank my co-chairs, Brian Junker and Lowell Taylor for their unending patience with meetings that were done by phone, fax, and email. Their insight and suggestions made this dissertation far better than it would have been without them. Besides being insightful, they are incredibly kind people with which to work. In addition, I wish to thank my committee members Steve Fienberg and Joel Greenhouse for their recommendations and advice. The many faculty in the Department of Statistics and the Heinz School of Public Policy and Management have given great support throughout my graduate career and they have made CMU a great place to attend school.

I also wish to thank the many people who housed me or my belongings in Pittsburgh once I was a traveling graduate student including Jason and Danielle Connor, Jeff and Carrie Palmer, and Beth Ayers. I was blessed to have many additional friends in both the statistics department and the Heinz school including Keith Hunter, Jeff Liebner, and Marnie Bertolet.

I also want to express my gratitude to the many colleagues at Temple University who offered advice, support, and friendship while I completed my dissertation including Kent McGuire, Heidi Ramirez, Will Jordan, Jim Byrnes, Lori Bailey, and Jennifer Cromley.

I thank Rachel Aucott, Emily Travers, Claudia Merritt, and Madeline Schofield for watching my daughter Abbey whenever I asked so I could work uninterrupted.

I am also blessed with a family who has provided more love and support than I deserve. In particular, I want to thank my husband Tim who edited countless pages and is always a shoulder to lean on and an ear for listening. I want to thank my sister Kristin Steuerle...
for her encouragement. I want to thank my daughter Abigail Lang Schofield for providing me with plenty of down time and a constant reminder of what is really important. Most especially I want to thank my father Eugene Steuerle for not only providing support, but also for editing many drafts of various chapters, sending along articles he thought I might be interested in, and constantly reminding me that I could do it. I want to thank my mother Norma Lang Steuerle who was the first one who suggested that I attend graduate school and who promised to help me every step of the way. She managed to do that despite passing away before I was even accepted.

Finally, I want to thank my grandfather Norman Lang who started the family’s relationship with CMU by taking a second job working as a housekeeper over 50 years ago. His own sacrifices of dropping out of college to support his wife and children—my grandmother and mother—made this all possible.
# Contents

Abstract  iii  

Acknowledgements  v  

1 Introduction  1  

2 Item Response Theory  7  

2.1 Classical Test Theory  7  

2.2 Basic Item Response Theory Models  8  

2.2.1 The 3-PL, 2-PL, and 1-PL Models  9  

2.3 Assumptions Used in Item Response Theory  11  

2.3.1 Item Response Functions  11  

2.3.2 Unidimensionality  11  

2.3.3 Local Independence  12  

2.4 Estimating IRT Models  13  

2.4.1 The Joint Maximum Likelihood Method  13  

2.4.2 The Marginal Maximum Likelihood Method  14  

2.4.3 Markov Chain Monte Carlo Methods  15  

2.5 Standard Error of $\hat{\theta}_j$  16  

2.6 Differences Between IRT and CTT  17  

2.7 Marginal Estimation Procedures  18  

2.7.1 Calculating the Integral  19  

2.7.2 The Conditioning Model  20
3 Current Methods

3.1 Current Methods ................................................. 26
   3.1.1 Elementary Linear Regression ......................... 26
   3.1.2 Errors In Variables .................................... 30
   3.1.3 Marginal Estimation Procedures ....................... 32

3.2 Three Simulation Studies ........................................ 36
   3.2.1 Simulations Using Elementary Linear Regression ........ 36
   3.2.2 Simulations Using Errors-in-Variables Model ........... 39
   3.2.3 Simulations Using Plausible Values Methodology ........ 43

3.3 Conclusion .................................................... 46

4 The MESE Model .................................................. 49

4.1 The Mixed Effects Structural Equations Model .................. 50
   4.1.1 The MESE Model as a Structural Equations Model ....... 52
   4.1.2 Identifiability of the MESE Model ...................... 53

4.2 Estimates in MESE .............................................. 56
   4.2.1 The Mean and Variance of the Posterior Distribution of $\beta_2$ 61

4.3 Bias in the Race Coefficient ................................... 63
   4.3.1 The Conditioning Model Revisited ..................... 63
   4.3.2 Measurement Error ...................................... 68

4.4 Assumptions in the MESE model ................................. 72
   4.4.1 The Prior Distribution on $\theta$ ....................... 72
   4.4.2 The IRT Model .......................................... 75

4.5 Recommended Test Length and Item Variability ................ 79

4.6 Comparing the MESE Model to Other Methods .................. 83
   4.6.1 MESE Model Compared to ELR .......................... 83
   4.6.2 MESE Model Compared to EIV .......................... 85
   4.6.3 MESE Compared to ME ................................. 88

5 Using the MESE to Analyze Black-White Wage Gaps ................. 91

5.1 Current Literature on Black White Wage Gaps ................. 91
CONTENTS

5.1.1 Black-White Wage Gaps ........................................ 91
5.1.2 Measuring Human Capital ...................................... 92
5.2 The 1992 NALS ...................................................... 96
5.3 Empirical Analysis of Black-White Wage Differences ............... 97
  5.3.1 Following Neal and Johnson ................................... 98
  5.3.2 Education and Test Scores ..................................... 101
  5.3.3 Interaction Effects ............................................. 105
  5.3.4 Residual Analysis .............................................. 108
5.4 Discussion .................................................................. 109
5.5 Appendix ............................................................... 118

6 Discussion and Future Work ............................................ 129
  6.1 Summary and Contributions ....................................... 129
  6.2 Future Work ........................................................ 132

References ................................................................. 137
List of Figures

2.1 Three Typical IRFs ......................................................... 9

4.1 Two IRFs with Different Discrimination Parameters ................... 78

4.2 Boxplots of the Theta Residuals for the Four Simulations where IRT Model Misspecification Occurred. ........................................... 79

4.3 Median Standard Errors for Different Test Lengths and Difficulty Parameter Variability Across Theta. ........................................... 81

5.1 Diagnostic Plots–MESE model for Married Men ....................... 119

5.2 Diagnostic Plots–MESE model for Married Men with an Education Variable 120

5.3 Diagnostic Plots–MESE model for Married Men with an Interaction Term . 121

5.4 Diagnostic Plots–MESE model for Single Men ............................. 122

5.5 Diagnostic Plots–MESE model for Single Men with an Education Variable . 123

5.6 Diagnostic Plots–MESE model for Single Men with an Interaction Term . 124

5.7 Diagnostic Plots–MESE model for Single Women .......................... 125

5.8 Diagnostic Plots–MESE model for Single Women with an Education Variable 126

5.9 Diagnostic Plots–MESE model for Single Women with an Interaction Term 127
List of Tables

3.1 Results from Simulations-Bias in Elementary Linear Regression Coefficients when Measurement Error is Not Modeled ........................................ 38
3.2 Results from Simulations-Typical Errors-in-Variables Models ............... 42
3.3 Results from Simulations-Various Methods Using Plausible Values as a Measure of Cognitive Skills ..................................................... 45
4.1 Results from Simulations-The Conditioning Model in ELR and MESE models 68
4.2 Results from Simulations-Increasing Measurement Error .................. 71
4.3 Results from Simulations-Testing the Shape of the Prior Distribution on \( \theta \) ......................................................... 74
4.4 Results from Simulations-Testing Misspecification of the IRT Model ........ 77
4.5 Selected Results from Simulations-Test Length and Variability of Items .. 82
4.6 Results from Simulations-ELR versus MESE ................................ 84
4.7 Results from Simulations-EIV versus MESE ................................ 87
4.8 Results from Simulations-PV versus MESE ................................ 89
5.1 Sample Characteristics of the NALS ............................................. 112
5.2 Baseline Regressions Using Elementary Linear Regression .................. 113
5.3 Regressions Comparing Med PV, PV, and 2 Specifications of the MESE Model114
5.4 Results from Simulations-ELR and MESE Models with Education and \( \theta \) .. 115
5.5 Regressions Comparing 3 Specifications of the MESE Model–Including just \( \theta \), including \( \theta \) and Ed and Including and Interaction between \( \theta \) and Race .. 116
5.6 Results from Simulations-ELR and MESE Models when There is an Interaction Term Between Test Score and Race ................................. 117
5.7 BIB Design of the 26 Booklets Used in the 1992 NALS . . . . . . . . . . . . 118
Chapter 1

Introduction

There are a number of studies within the social sciences that use cognitive test scores as independent variables in regression analyses. While there are some cases in which the predictive value of the cognitive test score itself is of interest, often the cognitive test score is being used as a proxy variable as a way to measure something else.

Measures of human capital\(^1\), ability, cognitive skills, IQ, etc.\(^2\) are important latent variables in social science analysis (see Hunt, 1995 for a discussion of the research on the role of intelligence in modern day society). Often, these analyses are regression-based and the measure of human capital is placed on the right-hand side of the regression equation as an independent variable,

\[
    w_j = \beta_0 + \beta_1 \theta_j + \beta_2 Y_j + \varepsilon_j
\]

\[
    \varepsilon_j \sim N(0, \sigma^2),
\]

where \(w_j\) is the dependent variable of interest, \(\theta_j\) is the latent human capital measure, and \(Y_j\) are the other covariates in the analysis. Since the ultimate goal of many of these analyses is to inform public policy, of great interest is the properties of these models.

Social scientists include these latent variables in their analysis for two reasons. First,

---

\(^1\)I use Deardorffs Glossary of International Economics definition of human capital as the knowledge and skill, that results from education, training, and experience, that makes an individual more productive.

\(^2\)I use the term human capital throughout, but for the non-economists, other terms including ability, cognitive skills, and IQ can be substituted.
social scientists are interested in the estimate of $\beta_1$ or the role human capital plays in predicting dependent variables. Political scientists interested in examining voting patterns use cognitive test scores as a descriptive demographic variable to characterize voters (see Venezky and Kaplan, 1998). Public health officials might want to understand how cognitive ability affects a patient’s understanding of and likelihood to follow doctor’s orders (for an example on the issue of health literacy in diabetes patients see Schillinger, et al, 2002; though the issue of health literacy is prevalent across many diseases and conditions).

Second, social scientists want to control for human capital in order to understand the role of other variables in the analysis, some of which may be highly correlated with human capital. Economists control for ability in order to better understand the role of parental financial resources in determining college participation rates (Dynarski, 1999). Psychologists often control for ability when trying to understand the role of motivation in learning situations (Huit, 2001). Economists control for human capital when studying black-white wage disparities in order to determine if the wage gaps are caused by skills gap or disparate treatment in the labor force. In these examples, the social scientist is less interested in the coefficient on ability and more interested in the estimate of $\beta_2$, the coefficient on other covariates in the regression equation. For example, in the case of understanding the role of parental resources on college participation rates, the main interest is on the estimate of the regression coefficient in front of the parental resources variable. Does having a parent with low resources affect an individual’s likelihood of going to college?

While the use of cognitive test scores as a measure of human capital is becoming more common, a more traditional approach uses a measure of education. (For an example using this approach see Trejo, 1997). Two different measures of education are quite common: the first is an indicator variable that denotes diplomas or certificates received and the second shows years of schooling. The use of these education variables to measure ability makes sense, since often the goal of education is to improve one’s ability, skills, and knowledge. However, educational attainment is a crude measure of ability. Measures of attainment that simply assess the number of years or achievement of an individual do not differentiate among the quality and strength of educational experiences. In the United States, in particular, the quality of education varies greatly across communities. Most educators agree that on
average schools in urban and rural communities have lower quality than schools in more affluent suburbs (U.S. Department of Education, 1996). In addition, measures of schooling do not quantify learning that takes place outside of a classroom. One way to overcome this crude measure of education is to collect much more data on the educational experience as in Black, Haviland, Sanders, and Taylor (2006).

In their landmark paper, Neal and Johnson (1996), who study black-white wage gaps contend that cognitive test scores such as the Armed Forces Qualifying Test (AFQT) are a better summary measure of pre-market human capital. Cognitive test scores show differences in the quality of education and also account for any learning that has taken place outside of a formal classroom structure.

A further and even more recent approach is to include both a cognitive test score and an education variable as seen in Lang and Manove (2006), who also study black-white wage gaps. While some may argue that the two variables are highly correlated, Lang and Manove (2006) argue that education is a visible signal about skills to potential employers and so is needed in the investigation. Employers rarely know what a person’s cognitive ability is upon hiring, but will be able to “see” a person’s educational attainment.

Cognitive test scores, of course, come with their own set of problems. Black, Haviland, Sanders and Taylor (2006) discuss three important issues that arise when using cognitive test scores. The first is that many cognitive tests may not measure the exact area of ability (or intelligence, knowledge, or human capital) that is needed in a particular analysis. For example, Neal and Johnson (1996) use the AFQT as a control of ability in understanding racial wage gaps. The AFQT score comes from four subtests of the Armed Forces Vocational Aptitude Battery (ASVAB): Word Knowledge (WK), Paragraph Comprehension (PC), Arithmetic Reasoning (AR), and Mathematics Knowledge (MK). It can be quite easy to argue that many of the skills needed in a job are not included in the list above.

Black et al.’s second issue is that many cognitive test scores suffer from racial (and sometimes gender) bias. While psychometricians have developed some methods for handling this issue, the problem remains (see Hambleton, R. K. and Rogers, H. J., 1996 for more). This issue is particularly problematic when investigating areas like racial wage gaps, because test bias (particularly bias against minorities) can induce bias in the regression results.
CHAPTER 1. INTRODUCTION

The third and final issue suggests that cognitive test scores are imperfect measures themselves. It is this last issue that this dissertation will address.

Test scores are measured with error. No one test can exactly measure a person’s ability in any area. A well-abled person can have a bad day and mis-answer many items on a test, while a less-abled person can have a particularly lucky day and guess well. Some test items are written better than others and produce a more accurate assessment of a person’s cognitive ability. In addition, most tests cannot cover the full range of the construct being measured because the test would be impossibly long. In order to account for this issue, most test makers report both a point estimate of the true score as well as standard errors around that score. In most cases of social science research (economic and other areas), the test score is placed as a fixed independent variable in the regression analysis and the standard error is ignored (see Venezky and Kaplan, 1998, Tyler, Murnane, and Willett, 2000, Raudenbush and Kasim, 1998, and Neal and Johnson, 1996 as a few examples). Lessons from errors-in-variable analysis (Anderson, 1984) state that using a variable measured with error as an independent variable in an analysis results in coefficient estimates that will be biased. Furthermore, if the variable measured with error is correlated with other covariates in the analysis, those coefficient estimates will also be biased.

It is against this backdrop that I present this dissertation. The purpose is to address the issue of modeling the measurement error when using a cognitive test score as an independent variable. While there are many areas of social science using cognitive test scores as a measure of ability, this dissertation focuses on the substantive issue of black-white wage gaps. Here, economists control for human capital in order to obtain a more accurate measure of the wage gaps. In addition, by controlling for human capital, economists can understand better why the wage gaps occurs. Is it a result of pre-market factors or possible discrimination in the workforce? In this dissertation, I demonstrate that choosing not to model the measurement error in the test score affects the resulting coefficients in black-white wage gap explorations. This can further affect the evidence used to answer the question of why the wage gap occurs. These biased results could have implications for whether policymakers use the evidence to focus time and money on reducing pre-market skills gaps or eliminating discrimination in the workforce.
Chapter 2 reviews two current methods in psychometrics for estimating ability using tests: Classic Test Theory (CTT) and Item Response Theory (IRT). In particular, I note the features, assumptions, fit, and estimation techniques of some of the models. The chapter also reviews marginal estimation techniques which are used when estimates of population parameters of ability are of interest, but individual estimates are poorly estimated. In addition, a brief comparison between IRT and CTT is provided. Chapter 3 discusses four current methods for modeling the measurement error of the latent independent variable in an analysis. The chapter notes the advantages and disadvantages of each of the current methods. This discussion leads to the introduction of a new model in Chapter 4: the Mixed Effect Structural Equations (MESE) model that deals with many of the disadvantages discussed with the current methods. Chapter 4 includes a set of small simulation studies that examine the many properties of the MESE model. Recommendations for the use of the MESE model in numerous situations are given. In addition, simulation studies compare the estimates of the MESE model to the estimates given by the current methods. Chapter 5 applies the MESE model in an examination of black-white wage gaps in married men, single men, and single women. I examine data from the 1992 National Adult Literacy Survey (NALS) and find that most of the wage gap can be explained by a skills gap. The results from the MESE model using the NALS data are compared to results using the current methods described in Chapter 3. I comment on the possible implications of not modeling the measurement error. In addition, this chapter gives some insight into the issue of including both education and the cognitive test score in the analysis, especially when the cognitive test score is measured with error and highly correlated to education. I also provide evidence that the return to skill is similar by race. Finally, Chapter 6 provides some concluding remarks as well as a discussion of possible future research.
CHAPTER 1. INTRODUCTION
Chapter 2

Item Response Theory

As noted in the previous chapter, observed test scores are often used as a measure of ability in a regression analysis. The purpose of this chapter is to review current methods for estimating ability in order to better understand the properties of the estimate that is used in the regression analysis. This chapter gives a short overview of classical test theory (CTT) and a longer overview of item response theory (IRT) noting the features of the models, the common assumptions of the models, and the techniques used for estimation and comparing fit of the models. For more see Stout, 2002 for a review of some of the psychometric research over the last 15 years. The chapter also describes marginal estimation procedures which estimate population parameters of ability that do not involve calculation of individual estimates.

2.1 Classical Test Theory: The Predecessor to Item Response Theory

In the 1950s, most educational measurement models were based on classical test theory (CTT), which allowed psychometricians to estimate a person’s ability or “true score” (Lord and Novick, 1968) based on her performance on a test. In CTT, the true score is defined as the expected value of the observed performance on the test of interest (Hambleton, Swaminathan, and Rogers, 1991),

\[ \phi_j = \theta_j + \nu_j, \]  

(2.1)
where $\phi_j$ is individual $j$'s observed test score, $\theta_j$ is the “true score” or latent ability trait, and $\nu_j$ is random error. The error is assumed to account for problems or issues that might differ in various test administrations (e.g., the health of the individual taking the test or the physical conditions of the room in which the test is being given).

Some common assumptions in CTT include:

\begin{equation*}
E(\nu_j) = 0
\end{equation*}
\begin{equation*}
\rho(\theta, \nu) = 0
\end{equation*}
\begin{equation*}
\rho(\nu_j, \nu_k) = 0,
\end{equation*}

where $j \neq k$. In these assumptions, $E(\nu_j)$ is the expectation of the standard error of measurement, $\rho(\theta, \nu)$ is the correlation between $\theta$ and $\nu$, and $\rho(\nu_j, \nu_k)$ is the correlation between two error random variables for two distinct measurements (Lord and Novick, 1968).

With these assumptions in place, it follows that in CTT,

\begin{equation*}
\theta_j = E(\phi_j),
\end{equation*}

for a given test taker where replication occurs in the test taker taking the same test over time and in a variety of situations. CTT was the dominant approach to estimating true scores until Lord and Novick (1968) replaced it with Item Response Theory (IRT).

### 2.2 Basic Item Response Theory Models

Item Response Theory (IRT) models are based on the concept that the set of traits underlying the performance on a test can be described by the item response function (IRF; also called the item characteristic curve). The IRF is typically a monotonically increasing function which specifies that as ability increases so does the probability of correctly answering a test item (Baker, 2001). Figure 2.1 shows three different IRFs. There are many different IRT models and they differ in three major ways: the shape of the IRF, the number of item parameters, and the number of examinee parameters. Below I describe one popular family of IRT models, the logistic models.
2.2. BASIC ITEM RESPONSE THEORY MODELS

2.2.1 The 3-PL, 2-PL, and 1-PL Models

The logistic family of models are unidimensional models (or models that assume that the test only measures one ability) that are best used with items that have dichotomous answers where no partial credit is possible.

The 3 parameter logistic (3-PL) model (Birnbaum, 1968) has the mathematical form:

\[
P(x_{ij} = 1 | \theta_j, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp(-1.7a_i(\theta_j - b_i))},
\]

where \(x_{ij}\) is the response of individual \(j\) to item \(i\) that is equal to 1 when the item is answered correctly and equal to 0 when the item is answered incorrectly. In the 3-PL model, the probability that individual \(j\) responds correctly to item \(i\) is conditioned on a cognitive ability \(\theta_j\), a “discrimination parameter” \(a_i\), a “difficulty parameter” \(b_i\), and a “guessing parameter” \(c_i\).

The discrimination parameter, \(a_i\) affects the slope of the curve, where a large \(a_i\) produces a steep slope and smaller \(a_i\) produces more gradual curves. Figure 2.1 shows three IRFs with three different discrimination parameters. IRT (3) (the red IRF) has the smallest discrimination parameter where \(a_3 = 0.5\). In IRF (2) (the black IRF), \(a_2 = 1\), and in IRF
(1) (the blue IRF), \( a_1 = 2 \). Notice how IRF (1) is the steepest. Thus as \( a_i \) increases, so too does an item’s ability to ‘discriminate’ among examinees of differing abilities. When \( \theta_j = b_i \), \( a_i \) is proportional to the slope of the IRF (Hambleton, Swaminathan, and Rogers, 1991). While technically \( a_i \) can be anywhere along the scale \((−\infty, \infty)\), more commonly \( a_i \) is defined on the range of \((0, 2)\) (Hambleton, Swaminathan, and Rogers, 1991).

The difficulty parameter, \( b_i \) affects the location of the IRF along the ability scale. Higher \( b_i \)s produce IRF’s with longer left-hand ‘tails’ and lower \( b_i \)s produce IRF’s with longer right-hand ‘tails’. As \( b_i \) increases, so too does an item’s difficulty—decreasing the probability that most examinees will answer the item correctly. When \( c_i = 0 \), \( b_i \) is equal to the value of \( \theta_j \) when the probability of a correct response is 0.5 (Hambleton, Swaminathan, and Rogers, 1991). In Figure 2.1, the three IRFs have three different difficulty parameters. In the case of IRF (3), \( b_3 = 1 \), while \( b_2 = 0 \) for IRF (2) and \( b_1 = −1 \) for IRF (1). Since IRF (1) has the smallest difficulty parameter, its location is the farthest to the left. The scale of the difficulty parameter tends to run from \((−2, 2)\) (Hambleton, Swaminathan, and Rogers, 1991).

The guessing parameter, \( c_i \) affects the location of the y-intercept (Hambleton, Swaminathan, and Rogers, 1991). Higher \( c_i \)s move the curve up on the y-axis with lower \( c_i \)s keeping the y-intercept closer to 0. In Figure 2.1, both IRF (2) and (3) have guessing parameters that are 0, while IRF (1) has a guessing parameter that is 0.2. Thus, the y-intercept of IRF (1) is higher. As \( c_i \) increases, so too does the probability that an item can be answered correctly through guessing.

When \( c_i = 0 \), the 3-PL model becomes the two parameter logistic (2-PL) model (Lord, 1952) with the following form

\[
P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1 + e^{x(-1.7a_i(\theta_j - b_i))}}.
\]

A further restricted form of the 3-PL model is the one parameter logistic (1-PL) also known as the Rasch model (Rasch, 1960). It has the following form:

\[
P(x_{ij} = 1|\theta_j, b_i) = \frac{1}{1 + e^{x(\theta_j - b_i)}}.
\]
2.3. ASSUMPTIONS USED IN ITEM RESPONSE THEORY

The IRF’s of different 1-PL models differ only in their location and not in their shape as the discrimination parameter is equal for all items in the Rasch model.

2.3 Assumptions Used in Item Response Theory

In order to use IRT models, it is important to understand the underlying assumptions that are commonly associated with the models. Of course, the three models discussed in Section 2.2 assume different numbers and types of characteristics of the items. As noted before, the 1-PL or Rasch model assumes that individuals cannot guess answers and that all items are equally discriminating, but of differing difficulty. The 2-PL model assumes that some items are more discriminating than others, that some items are more difficult than others, but still assumes that correct item answers cannot be guessed. The 3-PL model assumes different difficulty levels, different discriminating properties, and that items can be guessed with some probability. Despite these differences in the models, there are three assumptions which are common to the family of logistic models described in Section 2.2: the item response function, unidimensionality, and local independence.

2.3.1 Item Response Functions

All IRT models assume that the IRF reflects the true relationship between ability and item responses (Hambleton, Swaminathan, and Rogers, 1991). In the models described in Section 2.2, this means that the relationship is assumed to be logistic. Other IRT models assume the relationship to be normal ogive in shape (Lord, 1952). There are also non-parametric IRT models (Sijtsma and Molenaar, 2002) whose only assumption about the shape of the IRF is that it is non-decreasing. The assumption of monotonicity can be replaced with other smoothness assumptions as in Ramsey, 1991. More generally, see Sijtsma and Junker (2001) for an overview of non-parametric models.

2.3.2 Unidimensionality

The logistic family of IRT models assumes that there is only one ability or construct measured by the items on the test. Hambleton, Swaminathan, and Rogers (1991) point out that
this assumption is unlikely to be true as there are many factors including motivation, test anxiety, and general cognitive skills other than the one being measured by the test that can influence how well an individual answers the items. Certainly, there is an entire literature in the educational field that examines the effect of motivation of learning (see Huitt, 2001 for more).

Psychometricians often argue that unidimensionality can be assumed when a dominant factor influences test performance, even knowing that motivation and test anxiety can influence performance as well (Hambleton, Swaminathan, and Rogers, 1991). Multidimensional models have been proposed where two or more constructs of ability are modeled (see Adams, Wilson, and Wang 1997 as an example). In many cases, however, estimation problems exist for these models (Harrison, 1986).

2.3.3 Local Independence

Local independence assumes that conditional on ability, answers to items on a test are statistically independent of one another (Hambleton, Swaminathan, and Rogers, 1991). Lord and Novick (1968) define local independence in any population $P$, with fixed ability $\theta_j$, item responses $x_i$, and joint distribution $F$ as

$$F(x_{1j}, x_{2j}, ..., x_{mj} | \theta_j) = F(x_{1j} | \theta_j) F(x_{2j} | \theta_j) ... F(x_{nj} | \theta_j).$$

Local independence can also be assumed when more than one factor influences the responses to a test, as long as all ability factors are taken into account. For example, individuals who do not speak English may have difficulty with items on a math test written in English. If, however, all examinees speak, read, and write English equally well, then local independence will hold.

The assumption of local independence is important in making estimation of IRT models much easier.
2.4 Estimating IRT Models

When local independence is assumed, the joint probability of the responses to the whole test (the response vector) is equal to the product of the probabilities of each individual response. The IRT likelihood is

$$L(x_{1,1}, \ldots, x_{m,n}|\theta_1, \theta_2, \ldots, \theta_n) = \prod_j \prod_i P(x_{ij}|\theta_j)^{x_{ij}} (1 - P(x_{ij}|\theta_j))^{1-x_{ij}}. \quad (2.7)$$

In the case of the 3-PL model, there are $3M + N$ parameters to estimate (three item parameters for each of $M$ items and one ability parameter for each of the $N$ respondents).

IRT models clearly suffer from the problem of indeterminacy. In order to solve this problem, the values are set on an arbitrary scale – usually where $\mu_\theta = 0$ and $\sigma^2_\theta = 1$ where $\mu_\theta$ and $\sigma^2_\theta$ are the mean and variance of the ability distribution respectively (Hambleton, Swaminathan, and Rogers, 1991). This scaling will need to be taken into account when $\theta_j$s from different tests are compared.

Three commonly-used methods for estimating the IRT model are joint maximum likelihood methods, marginal maximum likelihood methods, and Markov Chain Monte Carlo methods. (For further information about estimating IRT models see Baker and Kim, 2004.)

2.4.1 The Joint Maximum Likelihood Method

The joint maximum likelihood method (Hambleton, Swaminathan, and Rogers, 1991) estimates the item parameters and ability at the same time. The likelihood which is maximized is

$$L(X_{ij} = x_{ij}|\theta_1, \ldots, \theta_N, \gamma_1, \ldots, \gamma_M) = \prod_j \prod_i P(X_{ij} = x_{ij}|\theta_j, \gamma_i), \quad (2.8)$$

where $x_{ij}$ is person $j$’s response to item $i$, $\theta_j$ is the true ability of person $j$, and $\gamma_i$ is the vector of item parameters ($a_i$, $b_i$, and $c_i$ as described above). Values of both the item parameters and the ability estimates that maximize the likelihood are determined using optimization methods such as the Newton-Raphson method or an E-M algorithm. In an E-M algorithm, initial ability estimates are chosen and treated as known. Item parameters are then estimated using maximum likelihood procedures. Then, these item parameter
estimates are taken as fixed, and new ability estimates are determined. These two steps are repeated until the two sets of estimates both converge. Standard errors are determined using the information function as usual (which I describe further in Section 2.5). The joint maximum likelihood method is implemented in the LOGIST (Wingersky, 1992) computer program for the 1-PL, 2-PL, and 3-PL models.

The joint maximum likelihood method cannot be used in the cases where an individual answered all items correctly or answered all items incorrectly, because the MLE of \( \theta_j \) will be either \(-\infty\) or \(\infty\) in such cases. It also cannot be used in the cases where an item was answered correctly by every examinee or where an item was answered incorrectly by every examinee, because the MLE of the item parameters will not exist. In addition, there are some response patterns (one example is where an examinee answers many difficult questions correctly and many easy questions incorrectly) where the method will fail to converge, because no finite absolute maximum will exist. In addition, the method does not produce consistent estimates in most cases (Haberman, 1977), though Douglas (1997) showed that if the number of items and the number of examinees tend to infinity together at controlled rates, consistent estimates are possible.

2.4.2 The Marginal Maximum Likelihood Method

The marginal maximum likelihood (MML) method (Bock and Aitkin, 1981) can be used to solve the consistency problem present in the joint maximum likelihood method. In the standard MML practice, the ability estimates are treated as missing data for all examinees. Next, \( \theta_j \) is assigned an underlying probability distribution and integrated out of the joint likelihood function

\[
L(X = x | \gamma_1, ..., \gamma_M) = \int_{\theta_1} \cdots \int_{\theta_N} \prod_j \prod_i P(X_{ij} = x_{ij} | \theta_j, \gamma_i) d\theta_1 ... d\theta_N. \tag{2.9}
\]

Marginal maximum likelihood estimates of the item parameters are then possible to obtain. These parameter estimates are consistent, because they are estimated without being conditional on the ability parameters. Once estimates of the item parameters are obtained, they are taken to be fixed and known so that estimation of \( \theta_j \) can proceed.
2.4. ESTIMATING IRT MODELS

In most cases, $\theta_j$ is estimated using maximum likelihood estimation, though Bayes estimates are also used. The MML method is similar to Empirical Bayes (EB) procedures (Carlin and Louis, 2000). In EB, the posterior distribution, $p(\theta_1, \ldots, \theta_N | X_{1,1}, \ldots, X_{M,N}, \gamma_1, \ldots, \gamma_M)$ is estimated using $p(\theta_1, \ldots, \theta_N | X_{1,1}, \ldots, X_{M,N}, \hat{\gamma}_1, \ldots, \hat{\gamma}_M)$ where $\gamma_i$ is estimated using the marginal distribution $p(X_{1,1}, \ldots, X_{M,N} | \gamma_1, \ldots, \gamma_M)$ by $\hat{\gamma}_i \equiv \hat{\gamma}_i(X)$. Standard errors of the $\theta_j$ estimates are determined using the information function (for maximum likelihood estimates) or the posterior distribution (when Bayes estimates are used). The MML method is implemented in the BILOG-MG (Zimowski et al., 2003) computer program.

MML methods work well in most cases, however, as Patz and Junker (1999) point out there are two drawbacks to the MML method. First, as the model becomes more complex, the E-M algorithm becomes harder to apply as a maximization method. In addition, standard error estimates of $\theta_j$ will be too small, because as is typical in EB procedures, the standard errors of in the estimated item parameters are not included.

2.4.3 Markov Chain Monte Carlo Methods

In order to handle both the complexity of the algorithms used to maximize the likelihoods and the uncertainty problem, Patz and Junker (1999) developed a method for estimation that jointly estimates the item and ability parameters using Markov Chain Monte Carlo (MCMC) methods

$$
\int_{\theta_1} \ldots \int_{\theta_N} \int_{\gamma_1} \ldots \int_{\gamma_M} P(X_{ij} = x_{ij} | \theta_j, \gamma_i) \times \pi(\theta_j) \pi(\gamma_i) d\theta_1 \ldots d\theta_N d\gamma_1 \ldots d\gamma_M. \tag{2.10}
$$

In this method, prior distributions ($\pi(\theta_j)$ and $\pi(\gamma_i)$) must be chosen for each of the parameters. Often, this method can be implemented in a high-level statistical language like R or S-Plus.

In the case where both item and ability estimates are wanted, then consistent estimates will occur only with both the number of items and the number of examinees tends to infinity. However, in cases where ability estimates are all that is desired, the MCMC algorithm can be used to produce consistent estimates when the number of items goes toward infinity. The Bayesian method is a bit slower in producing results, and unlike the
joint maximum likelihood method and the MML method where standard software exists to produce estimates, the researcher has to have enough statistical ‘know-how’ to program the MCMC method.

2.5 Standard Error of \( \hat{\theta}_j \)

While there are many ways to estimate \( \theta_j \) using the IRT models, none of the estimation techniques produce perfect estimates of \( \theta_j \). Measurement error exists. Fortunately, there are methods for estimating the standard error of \( \hat{\theta}_j \).

In cases where maximum likelihood estimates (MLEs) exist, the standard error of \( \hat{\theta}_j \) is given as

\[
SE(\hat{\theta}_j) = \frac{1}{\sqrt{I(\theta_j)}},
\]

(2.11)

where \( I(\theta_j) \) is the information function which is defined as the sum over \( i \) of \( I_i(\theta_j) \),

\[
I_i(\theta_j) = \frac{P'(x_i = 1|\theta_j)^2}{P(x_i = 1|\theta_j)(1 - P(x_i = 1|\theta_j))},
\]

(2.12)

where \( I_i(\theta_j) \) is the information provided by item \( i \) at \( \theta_j \), \( P(x_i = 1|\theta_j) \) is the item response function for item \( i \), and \( P'(x_i = 1|\theta_j) \) is the derivative of \( P(x_i = 1|\theta_j) \) with respect to \( \theta_j \). Since \( \theta_j \) is not known, \( I_i(\theta_j) \) is often estimated using \( I_i(\hat{\theta}_j) \).

Equation (2.12) translates into

\[
I_i(\theta_j) = \frac{a_i^2}{exp(1.7a_i(\theta_j - b_i))(1 - exp(-1.7a_i(\theta_j - b_i)))^2},
\]

(2.13)

for the 2PL model. Note that information is higher for an item with values of \( b \) closer to \( \theta_j \). Intuitively, this makes sense as it means that an item that is close in difficulty to a person’s true ability will give more information about the person’s ability than an item that is too easy or too hard. Information is also generally higher for items with high values of \( a_i \), suggesting that an item with a large ability to discriminate between two closely related abilities will provide more information.

The standard error of \( \hat{\theta}_j \) tends toward 0 as \( i \to \infty \). In addition, standard errors will be
lower for $\hat{\theta}_j$ as information increases. In IRT, the standard error of $\hat{\theta}_j$ varies for different $\theta_j$ even with the same test. IRT models the standard errors as heteroskedastic.

2.6 Differences Between IRT and CTT

The differences in the assumptions used in CTT versus the assumptions used in IRT have some implications for the estimates of $\theta_j$ that are produced.

First, estimates of $\theta_j$ generated from classical test theory are dependent on the test given. CTT models are unidentifiable as it is impossible to decompose the observed score into the true score component and the error component because $\theta_j$ and $\nu_j$ are both unknown. Methods have been developed to estimate the true score which tend to involve bounding of either the error or the true score. But the estimated true score in CTT remains dependent on these assumptions about the test error.

CTT then, is considered to be test oriented rather than item oriented (Hambleton, Swaminathan, and Rogers, 1991). Examination of items tends to happen in a “classic item analysis” that is often separate from the analysis of the test score itself. Classic item analysis includes looking at item p-values (or the probability of any randomly drawn examinee getting an item right), item-item and item-total correlations and measures of test reliability like Cronbach’s alpha (Cronbach, 1951). Finally, just as item analysis happens outside of the CTT model, so does test equating in the CTT framework. Test equating is not supported by the CTT model.

IRT, by comparison, is item oriented. The IRF includes information about the items in the model. In addition, test equating can be directly incorporated into the IRT model. Tests using different sets of items that all measure the same construct should produce similar ability estimates for individuals.

Finally, CTT assumes that the standard errors of the estimates of ability are homoskedastic by examinees under the assumption that $E(\epsilon) = 0$. However, this is unlikely to be true. As noted in Section 2.5, estimates of ability will be more or less precise depending on the number of items on the test, the match between the difficulty of the test and the ability of the individual, and the discrimination of the items. In IRT, the model assumes
heteroskedastic standard errors through the information function \( SE(\hat{\theta}_j) = \frac{1}{\sqrt{I(\theta_j)}} \) which is different for different values of \( \theta_j \).

2.7 Marginal Estimation Procedures: Estimating a Statistic that is a Function of \( \theta_j \)

Up until this point, this chapter has reviewed two sets of models (CTT and IRT) that use observed responses to a test as a way to estimate a respondent’s ability. I have discussed three estimation methods and noted a method to calculate the standard errors of the individual estimates of \( \theta_j \). Often, the estimates of \( \theta_j \) that come from the estimation techniques described in Section 2.4 are the estimates used as fixed, independent variables in the regression analyses described in Chapter 1. In many of these regression analyses, the standard errors of \( \theta_j \) are not modeled. I show in Chapter 3 how not modeling the error in \( \theta_j \) can lead to biased estimates of regression coefficients.

In applications like the regression models described in Chapter 1, point estimates of \( \theta_j \) are not sufficient and a distribution of \( \theta_j \) is needed. In these cases, the statistic of interest is not \( \theta_j \) but rather a function of \( \theta_j \). In the regression analysis example of Chapter 1 where \( \theta_j \) is used as an independent variable in a regression analysis, the statistic of interest is the coefficient on \( \theta_j \) in the linear model. Mislevy (1991) has shown that treating the estimated \( \theta_j \)'s as fixed and known can lead to substantial bias in population parameter estimates. In the case of a regression analysis, the bias in the regression coefficient could cause problems if it led the researcher to make incorrect inferences. A better approach that corrects bias in population parameter estimates is marginal estimation procedures (Mislevy, 1991) where population statistics are measured directly from the item responses and examinee background characteristics without estimating individual level test scores. I now review this process.

Let \( s(\theta, X, Y) \) be the statistic of interest. In Chapters 3 and 4, for instance, I am interested in estimating the black-white wage gap. A common analysis of the wage gap includes a regression equation where log wages, \( w_j \) is estimated to be a function of demographic variables, \( Y_j \), (for instance education, age, and potential experience in the workforce), item
responses, $X_{ij}$, which come from a test of cognitive ability, and $\theta_j$, a latent ability construct. In such an analysis, the statistic of interest $s$ is the estimated regression coefficients on race which is dependent on $\theta$, $X$ and $Y$.

The conditional expectation of $s(\theta, X, Y)$ is

$$s(X, Y) = E[s(\theta, X, Y)|X, Y] = \int s(\theta, X, Y)p(\theta|X, Y, \alpha, \gamma)d\theta,$$

(2.14)

where $\alpha$ = parameters in the ‘hyperprior’ for $\theta$ and $\gamma$ = parameters in the likelihood for $X|\theta$ (individually called a, b and c in the sections above). Using Bayes Theorem

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

(2.15)

where $\kappa_{\alpha, \gamma}$ is a conditioning constant. The IRT model is encoded in $p(X|\theta, Y, \alpha, \gamma)$.

Equation (2.15) can be simplified. First, standard in IRT, assume responses to items on a test are only dependent on the latent ability and independent of one another. In addition, assume that $\theta$ is not dependent on $\gamma$ and each individual’s ability is independent of one another, leaving the following measurement model that is a multilevel/Bayes model of the form

$$\theta_j \sim p(\theta_j|\alpha, Y_j)$$

(2.16)

$$X_{ij} \sim p(X_{ij}|\theta_j, \gamma_i).$$

(2.17)

In practice, calculating the integral in (2.14) can be daunting. Two different methods have been developed which aid in the calculation of the integral: Plausible Value (PV) Methodology and Marginal Maximum Likelihood (MML) Procedures.

### 2.7.1 Calculating the Integral (2.14)

The Plausible Value (PV) Methodology uses Monte Carlo integration through the use of multiple imputations to calculate the integral while the Marginal Maximum Likelihood
(MML) procedures use numerical quadrature.

In Monte Carlo integration, random points over a sample domain are picked. The sample domain is defined as a superset of the domain which spans the integral that should be estimated. An easy and intuitive way in which to implement Monte Carlo integration, is through the use of multiple imputation. While multiple imputation is often considered a way in which to deal with missing data, 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 responses to give a data set with no missing values so that $s$ can then be evaluated. Because approximations of the standard errors of $s$ will be underestimated if only a single imputation is used, Rubin (1977) suggested using multiple draws (Rubin suggests five draws though any number could be drawn). 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. Thus, the PV methodology requires a primary analyst to draw the multiple imputations (called plausible values in PV methodology) to be included in the data set for use by secondary analyst.

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 $\theta$ is large or $s(\theta, X, Y)p(\theta|X, Y, \alpha, \gamma)$ is not smooth, estimates using numerical quadrature are more problematic.

2.7.2 The Conditioning Model

In addition to calculating the integral, marginal estimation procedures also must be concerned with what to include in the conditioning model (2.16).

Consider a situation in which the posterior for $\theta$ depends on a different set of covariates $Y^*$ than the covariates $Y$ in $s(\theta, X, Y)$.

$$s(X, Y, Y^*) = \int s(\theta, X, Y)p(\theta|X, Y^*, \alpha, \gamma)d\theta.$$ (2.18)

Biases will occur when $Y$ is not contained in $Y^*$, because $E[s(X, Y, Y^*)|X, Y]$ will not
necessarily be equal to \( s(X, Y) \). Thus, it makes sense to require at least \( Y \subseteq Y^* \). When \( Y \subseteq Y^* \), the multilevel model will lead to inconsistent estimates of the statistic of interest without ever requiring the calculation of individual estimates of \( \theta_j \). So, in the example above when the issue of black-white wage gaps is being examined, \( Y \) should include race and any other covariates that are in the model like age, gender, and possibly marital status.

The PV methodology uses a “saturated” conditioning model. Since the PV methodology produces the five plausible values to be included in a data set, the plausible values need to be conditioned on as many covariates as possible so that secondary analysts can use the same five plausible values regardless of what covariates are included in their model. There are cases, however, when additional data is collected after the calculation of the plausible values. In order to use this new data in the model, the plausible values will need to be recalculated.

The MML procedures use a small conditioning model. Advocates of the MML approach argue that when \( Y \) contains many variables, estimates of \( s(X, Y) \) will be computationally inefficient. 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 \( \theta \) is normal. In some populations, Dresher’s analysis demonstrates that the distribution of \( \theta \) is not normally distributed. She found a skewness of up to \(-0.3\) in some of the populations she studies. While it is true that the use of a small conditioning model is computationally more efficient, secondary analysts who use the MML procedures to calculate \( s(\theta, X, Y) \) will have to recalculate the integral for every analysis in order to change \( Y \) to include the appropriate background variables. There is a software program called AM (Cohen, 2002) which recomputes the integral using numerical quadrature for each analysis depending on what variables are included in the conditioning model. The AM software is currently unable to compute statistics when \( \theta \) is an independent variable in an analysis.

This chapter has briefly examined some of the issues involved in estimating ability using CTT or IRT. Methods for obtaining unbiased individual point estimates of ability and unbiased estimates of population parameters of \( \theta \) have been reviewed. In the following chapter, I examine some current methods that use the estimates of \( \theta_j \) from CTT or IRT as
independent variables in a linear regression model.
Chapter 3

Current Methods in Using $\theta_j$ as an Independent Variable

In Chapter 1, I introduced some areas in social science in which a measure of human capital is used as an independent variable in an analysis. One such specific area of study is the examination of black-white wage gaps. There is a large literature in economics where one of the main purposes is to determine if black and white people are treated differently in the labor market (see Altonji and Blank, 1999 for a review of this literature). Many labor economists use empirical evaluation (e.g., regression based approaches) to understand the gap and its underlying causes. One approach is to estimate a linear regression in which the dependent variable is the log wage and independent variables include an indicator variable for race (e.g., 1 if black and 0 if white) along with a series of demographic variables intended to “control for” differing levels of human capital that individuals bring to the labor market. The central idea is that if, after controlling for observable human capital differences, there remains an estimated race wage gap, one can infer that there is disparate treatment in the labor market.

One way to control for human capital in the regression analysis is to use an independent measure of the cognitive ability as a summary statistic for the skills a worker obtained in

\footnote{For the remainder of this chapter and the two chapters following, I focus on the issue of black-white wage gaps. However, in Chapter 6, I return to a broader discussion of the use of cognitive test scores as independent variables in other areas of social science as well.}
previous life experiences. A highly influential example of just such an analysis comes from
the work of Neal and Johnson (1996). Their analysis includes a regression similar to the
form

\[ w_j = \beta_0 + \beta_1 \theta_j + \beta_2 B_j + \varepsilon_j \]  
\[ \varepsilon_j \sim N(0, \sigma^2), \]  

where \( w_j \) is log wages, \( \theta_j \) is a measure of skill taken from the Armed Forces Qualifying Test
(AFQT), and \( B_j \) is a dummy variable that is 1 when the individual is black and 0 when the
individual is white. Neal and Johnson (1996) find that when their regression conditions
on a measure of cognitive skills that the estimated wage gap declines significantly from the
wage gap when there is no control for human capital. They conclude then that the wage
gap primarily reflects a skills gap instead of significant disparate treatment in the labor
market.

Neal and Johnson use the AFQT score as data in their analysis. In so doing, they
implicitly assume that their measure of skill, \( \theta_j \) is fixed and measured without error. In
Chapter 2, I note that test scores are measured with error and the observed test score
does not reflect the true score exactly. In addition, Section 2.7 discussed how population
parameters that are calculated using individual estimates will be biased. So, actually, Neal
and Johnson (1996) estimate the equation,

\[ w_j = \beta_0 + \beta_1 \phi_j + \beta_2 B_j + \varepsilon_j \]  
\[ \varepsilon_j \sim N(0, \sigma^2), \]  

where \( \phi_j \) is an observed test score measured with error. Errors-in-variables analysis shows
that bias will exist in the estimates of the regression coefficients in (3.2) unless the error in
\( \phi_j \) is modeled.

Work like that of Neal and Johnson (1996) is important because it provides evidence
about the causes of the racial wage gap. The regression coefficients help to determine

\(^2\text{Neal and Johnson (1996) also include age in their regression. Other analyses include other covariates; e.g., Lang and Manove (2006) include education and age.}\)
whether or not there is disparate treatment in the labor market for blacks or whether the wage gap is reflective of a skills gap. Thus, bias on the coefficients could lead to incorrect inferences on the part of the researcher. This could further lead to policies that focuses on eradicating disparate treatment in the work force when the skills gap is the primary cause of the wage gap or vise versa depending on the direction of the bias.

Analyses like those of Neal and Johnsons (1996) where an observed test score is used as a measure of $\theta_j$ without accounting for the measurement error are common across various areas of social science. As an example, Tyler, Murnane, and Willett (2000) examine the labor market signaling value of obtaining a General Education Development (GED) equivalency degree. Their analysis explores whether the increased wages one obtains after completing a GED is due to the signaling value of getting the degree or the increased skills one might obtain from studying for the exam. In their analysis, they include both an indicator variable for obtaining a GED as well as GED test scores which they include without error. Given the high correlation between obtaining a GED and GED scores, there may be bias in the regression coefficients on both sets of variables. Another example of work using an observed test score comes from political science. Using data from the 1992 National Adult Literacy Survey (NALS), Venezky and Kaplan (1998) develop a predictive model of voting behavior and find that when controlling for education and cognitive skills that cognitive skills does not play a particularly large role in predicting voter turnout. Once again, however, education and skills are highly correlated and when skills is modeled without error, bias will occur on both the estimates of the regression coefficients of skills and education.

It is against this backdrop that I examine how the measurement error in the observed test score that is used as a measure of $\theta_j$ when $\theta_j$ is an independent variable in a regression analysis affects the estimates of the regression coefficients. In this chapter, I examine four methods currently used in the literature for handling the measurement error problem: ignoring the measurement error using Elementary Linear Regression (ELR) as Neal and Johnson (1996) do, classic errors-in-variables (EIV) analysis, the Plausible Value (PV) Methodology, and marginal maximum likelihood (MML) methods. Each of these methods is described below.

Three small simulations studies demonstrate the drawbacks of each model. In elemen-
tary linear regression, large bias can result from ignoring the measurement error completely. While classic EIV methods will model the measurement error, EIV models are typically unidentifiable and are unable to model an error structure that is different for each individual. The PV methodology will accurately model the error in most situations, but can only be used by the typical social scientist when a primary analyst produces the plausible values. Since this is rarely done with commonly used tests and assessments, this method is not particularly universal. In addition, even for those tests and assessments for which plausible values are produced, there are some cases in which the conditioning model used to produce the PVs will not include covariates that are in the regression equation. Finally, the MML model has potential for solving the measurement error, but as of this writing, the software programs which implement the MML model do not allow for $\theta_j$ to be an independent variable. The chapter concludes with a brief discussion of the need for a model that models the measurement error, allows the magnitude of the error to differ from individual to individual, does not require a primary analyst to produce plausible values for use by the social scientist, and can be implemented with any test or assessment currently in use.

3.1 Current Methods

3.1.1 Elementary Linear Regression

In elementary linear regression analysis (Gauss, 1809) which is the model that Neal and Johnson (1996) use in their paper, a dependent variable is assumed to have a linear relationship with a function of some number of predictor variables. The coefficients in the model are often estimated using ordinary least squares estimation (OLS). Underlying the elementary linear regression model is the assumption that the dependent variable is subject to error. The error ($\varepsilon_j$) is assumed to be a random variable that follows a normal distribution with a mean of zero, and a constant variance ($\sigma^2$). The errors are additionally assumed to be uncorrelated such that $\text{cov}(\varepsilon_j, \varepsilon_k) = 0$ when $j \neq k$.

The independent variables are assumed to be linearly independent and error-free. Neal and Johnson (1996) in using model (3.1) assume that the test score perfectly models the true ability. The choice to model the observed test score as a fixed variable makes sense in
some situations. If the test is assumed to be fixed and unchanging, then the observed test score measures the respondent’s ability at the particular time of the test on that particular test.

If, instead the social scientist wants to either generalize across tests (for example, generalize across different versions of the SATs, or the AFQT) or wants the test construct to be viewed as a broader construct (e.g., Neal and Johnson using the AFQT to measure human capital) then the observed test score can no longer be viewed as a fixed, error-free observation. Rather, it is being used as a fallible measure of a latent variable \( \theta_j \) for which measurement error must be modeled.

Having said this, it might be fair to argue that even with perfect measures ability \( \theta_j \), this does not get around the fundamental problem that skills themselves are merely an imperfect proxy for human capital that is valued in the labor market. That is, one might think of the error as having two components: the error in using a test score as a measure of skills and the error in using skills as a proxy for human capital. Only the former error is dealt with when modeling the measurement error. While there may be no satisfactory way of dealing with the error in using skills as a proxy for human capital, it seems that modeling the measurement error in estimating \( \theta_j \) is clearly a helpful step in the right direction.

It is easy to see the extent of the bias in the regression coefficient estimates when ignoring the error. In this first example, assume that I know \( \theta_j \) exactly,

\[
\begin{align*}
  w_j &= \beta_0 + \beta_1 \theta_j + \beta_2 B_j + \varepsilon_j \\
  \varepsilon_j &\sim N(0, \sigma^2).
\end{align*}
\]  

Let \( \bar{w}_B \) and \( \bar{w}_W \) be the mean of observed log wage for blacks and whites respectively, and \( \bar{\theta}_B \) and \( \bar{\theta}_W \) be the corresponding means for the skills measure. Define the estimated covariances of \( \theta \) and \( w \) for the two demographic groups to be \( \hat{\sigma}_{\theta w, B} \) and \( \hat{\sigma}_{\theta w, W} \), and the estimated variance of \( \theta \) for the groups to be \( \hat{\sigma}_{\theta \theta, B} \) and \( \hat{\sigma}_{\theta \theta, W} \). Finally let \( \hat{\eta} \) be the fraction
of the sample that is black. The OLS estimators are
\[
\begin{pmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2
\end{pmatrix} = (Z^TZ)^{-1}Z^Tw,
\]
where \(Z\) is the data matrix and \(w\) is the response variable. Now with some algebra I can show that
\[
\begin{pmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2
\end{pmatrix} = \begin{pmatrix}
\hat{\eta}(1 - \hat{\eta}) & \hat{\eta}(1 - \hat{\eta})(\hat{\theta}_B - \hat{\theta}_W) \\
\hat{\eta}(1 - \hat{\eta})(\hat{\theta}_B - \hat{\theta}_W) & \hat{\eta}\hat{\sigma}_{\theta, B} + (1 - \hat{\eta})\hat{\sigma}_{\theta, W} + \hat{\eta}(1 - \hat{\eta})(\hat{\theta}_B - \hat{\theta}_W)^2
\end{pmatrix}^{-1} \times
\begin{pmatrix}
\hat{\eta}(1 - \hat{\eta})(w_B - w_W) \\
\hat{\eta}\hat{\sigma}_{\theta, B} + (1 - \hat{\eta})\hat{\sigma}_{\theta, W} + \hat{\eta}(1 - \hat{\eta})(\hat{\theta}_B - \hat{\theta}_W)(w_B - w_W)
\end{pmatrix},
\]
which when multiplied out gives OLS estimators of \(\beta_1\) and \(\beta_2\) that are, respectively,
\[
\hat{\beta}_1 = \frac{\hat{\eta}\hat{\sigma}_{\theta, B} + (1 - \hat{\eta})\hat{\sigma}_{\theta, W}}{\hat{\eta}\hat{\sigma}_{\theta, B} + (1 - \hat{\eta})\hat{\sigma}_{\theta, W}}, \quad (3.4)
\]
and
\[
\hat{\beta}_2 = (w_B - w_W) - \hat{\beta}_1(\hat{\theta}_B - \hat{\theta}_W). \quad (3.5)
\]
Under the assumptions that \(\theta_j\) and \(B_j\) are both measured without error, and \(\varepsilon\) is uncorrelated with these explanatory variables—the OLS estimators are of course consistent:
\[
\text{plim}(\hat{\beta}_1) = \frac{\eta\sigma_{\theta,w,B} + (1 - \eta)\sigma_{\theta,w,W}}{\eta\sigma_{\theta,B} + (1 - \eta)\sigma_{\theta,W}} = \beta_1, \quad (3.6)
\]
and
\[
\text{plim}(\hat{\beta}_2) = (\mu_{w,B} - \mu_{w,W}) - \beta_1(\mu_{\theta,B} - \mu_{\theta,W}) = \beta_2. \quad (3.7)
\]
This set-up allows me to easily examine what the consequences are for the estimators if cognitive skills \(\theta_j\) are poorly measured. In the simplest case the “true value” of skills, \(\theta_j\) for individual \(j\) is unobserved, and instead \(\phi_j\) which is contaminated with error \(\nu_j\) (with mean 0 and variance \(\sigma^2_{\nu,j}\)) is observed. I assume \(\phi_j\) to be independent of other variables in
3.1. CURRENT METHODS

the model. Then, \( \phi_j = \theta_j + \nu_j \). If I now estimate the regression coefficients without further modeling the error

\[
w_j = \beta_0^* + \beta_1^* \phi_j + \beta_2^* B_j + \varepsilon_j
\]

\( \varepsilon_j \sim N(0, \sigma^2) \).

I find

\[
\text{plim}(\hat{\beta}_1^*) = \frac{\eta \sigma_{\theta w, B} + (1 - \eta) \sigma_{\theta w, W}}{\eta \sigma_{\theta \theta, B} + (1 - \eta) \sigma_{\theta \theta, W} + \sigma_{\nu \nu}}.
\]

Comparing this latter expression with (3.6), I immediately derive a result that is familiar in the errors-in-variable literature, \( |\text{plim}(\hat{\beta}_1^*)| < |\beta_1| \); the coefficient on the mismeasured variable is biased toward 0. In examining black-white wage gaps, I have strong theoretical reasons to believe \( \beta_1 \) is positive (I believe increasing skills should increase wages), so when cognitive skills are poorly measured, I expect to be underestimating the market returns to those skills.

More importantly, in this case, there will also be bias in the estimated race indicator coefficient:

\[
\text{plim}(\hat{\beta}_2^*) = (\mu_{\theta, B} - \mu_{\theta, W}) - \beta_1^* (\mu_{\theta, B} - \mu_{\theta, W}),
\]

where \( \beta_1^* < \beta_1 \). In the U.S. there is differential access to quality education, so I expect that on average blacks will have lower \( \theta_j \)'s than whites. Thus, \( (\mu_{\theta, B} - \mu_{\theta, W}) < 0 \), so \( \text{plim}(\hat{\beta}_2^*) < \beta_2 \). I will be overestimating the wage gap. This could become problematic if I used this analysis as evidence of labor market discrimination when in fact there was none.

The problem with choosing to ignore the measurement error in an independent variable in a regression is well documented in the literature of errors-in-variable analysis (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. Coefficients on other variables can also be biased. Using arguments related to the one above, Bollinger (2003) notes that test scores are measured with error and that biases will occur in the regression coefficients. In his analysis of black-white wage gaps, he uses a typical errors-in-variables
model to handle the measurement error issue.

### 3.1.2 Errors In Variables

Errors-in-variables (EIV; Anderson, 1984) models assume that one or more of the independent variables has measurement error. When an EIV model is used instead of an elementary linear regression analysis, (3.1) changes to become:

\[
\begin{align*}
    w_j &= \beta_0 + \beta_1 \theta_j + \beta_2 B_j + \varepsilon_j \\
    \phi_j &= \delta + \Gamma \theta_j + \nu_j
\end{align*}
\]  

(3.9)  

(3.10)

where \( \phi_j \) is the observed test score (observed with error), \( \theta_j \) is the true unobserved ability, \( \delta \) and \( \Gamma \) are coefficients in the model of the measurement error, \( \varepsilon_j \) is the regression error which is often assumed to be \( N(0, \sigma_{\varepsilon\varepsilon}^2) \), and \( \nu_j \) is the measurement error of the observed test score \( \phi_j \) which is often assumed to have mean 0 and variance \( \sigma_{\nu\nu}^2 \).

In the EIV model, the observed test score is no longer assumed to be fixed without error. Rather, it is assumed that \( \phi_j \) is observed and \( \theta_j \) and \( \nu_j \) are unobserved. The EIV model uses CTT in its modeling of the observed test score. Note that (3.10) looks very similar to (2.1) from Chapter 2.

Like CTT models, EIV models, are often unidentifiable. Stefanski (1990) notes that in order to be identifiable, one of three things is needed, 1) additional data, (e.g. replicate measurements of \( \phi_j \)), 2) the distribution of \( \theta \), or 3) additional information in the form of distributional or moment restrictions on either or both of the error distributions. Cheng and Van Ness (1999) document six assumptions about one or both of the error distributions that will make the model identifiable: knowing \( \sigma_{\varepsilon\varepsilon}^2 \), knowing \( \sigma_{\nu\nu}^2 \), knowing the ratio \( \sigma_{\varepsilon\varepsilon}^2 / \sigma_{\nu\nu}^2 \), knowing both \( \sigma_{\varepsilon\varepsilon}^2 \) and \( \sigma_{\nu\nu}^2 \), knowing \( \beta_0 \), or knowing the reliability ratio \( \kappa \), where \( \kappa = \frac{\sigma_{\varepsilon\varepsilon}^2}{\sigma_{\varepsilon\varepsilon}^2 + \sigma_{\nu\nu}^2} \) (Cheng and Van Ness, 1999). In addition, Klepper and Leamer (1984) provide a method for estimating bounds of the regression coefficients even when none of the six assumptions above are known. The lower bound is the coefficient when using a direct regression or when...
3.1. CURRENT METHODS

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.

In his study of black-white wage gaps, Bollinger (2003) follows Klepper and Leamer (1984) to estimate bounds for the coefficients on race and AFQT score. His estimated bounds are quite large. He finds that the coefficient on race for men is anywhere in the range (-0.07, 1.26) and for women is anywhere in the range (0.04, 1.39). Since Bollinger is examining log wages, this translates into a range for men of blacks making approximately 7% less than whites to blacks making 126% more than whites when controlling for human capital. Bollinger (2003) is able to conclude that the estimated black-white wage gap likely is overestimated when an elementary linear regression is used instead of a model that accounts for the error. However, without further information or assumptions, point estimates of the coefficients are not possible with the classic errors-in-variables model using Klepper and Leamer bounds.

More importantly for this discussion, by assuming a CTT model in (3.10), EIV models assume that the magnitude of the error is the same for all individuals in the population. Recall from Chapter 2 that CTT postulates that the observed score is a linear function of the true score and random error (Lord and Novick, 1968). However, IRT suggests that the standard error of \( \hat{\theta}_j \) is likely different for different \( \theta_j \). Recall from Section 2.5 that the standard error of \( \hat{\theta}_j \) is lower when the test has items with difficulty parameters closer to \( \theta_j \) and/or when the items on the test have high values of discrimination parameters. Thus, when the population of people all have similar \( \theta_j \)s or when the test is sufficiently long to have many items across the distribution of \( \theta \), the assumption that the magnitude of the error is likely not a problem. However, when the test focuses on only one small span of \( \theta \), the magnitude of the error by individual could vary greatly. Later in the chapter, I demonstrate through the use of simulation, some possible problems when the assumption that the error is the same across the population is used.

Because of the two problems with EIV models, a model that uses IRT instead of CTT to estimate \( \theta_j \) is likely to be desired.
3.1.3 Marginal Estimation Procedures: Plausible Value Methodology and Marginal Maximum Likelihood Methods

Chapter 2 described two marginal estimation methods for estimating population parameters of $\theta$: Plausible Value Methodology (PV) and Marginal Maximum Likelihood (MML) methods. PV methodology and MML methods can be used when $\theta_j$ is an independent variable in a regression as well. If I specify the multilevel model in (2.16)-(2.17) to the purposes of examining black-white wage gaps, I have

$$
\theta_j \sim p(\theta_j|B_j) \quad (3.11)
$$

$$
X_{ij} \sim p(X_{ij}|\theta_j, \gamma_i),
$$

where $\theta_j$ is the true ability of individual $j$, $B_j$ is an indicator variable that is 1 if the respondent is black and 0 if the respondent is white, $X_{ij}$ are the item responses for individual $j$ to item $i$, and $\gamma_i$ are the item parameters, noted $a_i$, $b_i$ and $c_i$ throughout Chapter 2. The model is described in more detail in Chapter 2, but it is important to note a few of the assumptions inherent in this multilevel model.

First, the IRT model is encoded in the likelihood $p(X_{ij}|\theta_j, \gamma_i)$. This implies that the marginal estimation procedures recognize that the standard errors for different $\theta_j$ could be different, unlike the EIV model.

Second and somewhat counterintuitive, the prior on $\theta_j$ must be conditioned on race in order to avoid bias. The marginal estimation procedures rely on the concept that it is possible to estimate the conditional expectation of $s(\theta, X, Y)$ using $s(X, Y) = E[s(\theta, X, Y)|X, Y]$. Section 2.7.2 demonstrated that if the posterior for $\theta$ depends on a different set of covariates $Y^*$, $E[s(X, Y, Y^*)|X, Y]$ will not necessarily be equal to $s(X, Y)$. In practice, this means that $Y$ must include race, because the population statistic of interest, the regression coefficient on race, includes race. In addition to race, $Y$ must also include any additional covariates that are in the regression equation. In examining black-white wage gaps, this can include such variables as marital status, education, and age.
3.1. CURRENT METHODS

Plausible Value Methodology When $\theta_j$ is an Independent Variable

When the PV methodology is used to estimate $s(\theta, X, Y)$ in the multilevel model, estimates of $s$ can be calculated based on the average of the $Q$ estimates where $Q$ is the number of plausible values,

$$S_Q = \frac{1}{Q} \sum_{q=1}^{Q} s(q) \equiv \frac{1}{Q} \sum_{q=1}^{Q} s(\theta^{(q)}, X, Y). \quad (3.12)$$

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

$$V_Q = U_Q + \left(1 + \frac{1}{Q}\right) D_Q, \quad (3.13)$$

where

$$U_Q = \frac{1}{Q} \sum_{q=1}^{Q} u(q) \equiv \frac{1}{Q} \sum_{q=1}^{Q} u(\theta^{(q)}, X, Y); \text{ and}$$

$$D_Q = \frac{1}{Q - 1} \sum_{q=1}^{Q} (s(q) - S_Q)^2.$$

The estimate (3.13) incorporates both model-based uncertainty (in $U_Q$) and Monte-Carlo uncertainty (in $D_Q$).

In the case of estimating the regression coefficients in an analysis of black-white wage gaps, the estimates of the regression coefficients are calculated as the average of the estimated coefficients found by analyzing the regression five different times using each of the plausible values $\theta^{(q)}$ as the measure of $\theta_j$. Unfortunately, there are many studies in the literature which do not follow these suggested guidelines for estimating $s$. In particular, researchers use the first plausible value or the median of the five plausible values as an observed test score in their analysis. (For some examples, see Blau and Kahn, 2000 and Green and Riddell, 2003.)

The use of the first plausible value as the measure of $\theta_j$ will mainly produce problems in the calculation of the standard errors. Using only one plausible value will result in standard errors that do not include the Monte-Carlo uncertainty ($D_Q$ in (3.13)) and are underestimated. The larger problem, though, comes when researchers use the median plausible...
value as a fixed estimate of $\theta_j$ in their analysis. If every respondent’s median PV is used, Monte Carlo draws from the tails of the distribution will rarely be included in the regression analysis. As a result, the distribution of $\theta$ will be thinner, producing an estimate of $\theta_j$ that is pulled toward the group means.

Recall the linear regression model in (3.3). I showed that the OLS estimators in this model are consistent when there is no error in estimating $\theta_j$. I have argued that using the median plausible value will pull the values of the estimated $\theta_j$s toward the group means. Suppose $\theta_j^{**}$ is a measure of $\theta_j$ that contains values pulled toward group means, e.g., in the simplest case, for some $0 < \alpha < 1$,

$$\theta_j^{**} = \alpha \theta_j + (1 - \alpha)[B_j \mu_{\theta B} + (1 - B_j) \mu_{\theta W}],$$  
(3.14)

for individual $j$ where $\mu_{\theta B}$ and $\mu_{\theta W}$ are the true mean of $\theta$ for blacks and whites respectively. If these data are used in elementary regression equation (3.1), it is easy to show that the resulting estimators will have these properties:

$$\text{plim}(\hat{\beta}_1^{**}) = \frac{\eta \alpha \sigma_{\theta y, B} + (1 - \eta) \alpha \sigma_{\theta y, W}}{\eta \alpha^2 \sigma_{\theta y, B} + (1 - \eta) \alpha^2 \sigma_{\theta y, W}} = \frac{\beta_1}{\alpha} > \beta_1,$$  
(3.15)

and

$$\text{plim}(\hat{\beta}_2^{**}) = (\mu_{y, B} - \mu_{y, W}) - \frac{\beta_1}{\alpha} (\mu_{\theta B} - \mu_{\theta W}) > \beta_2,$$  
(3.16)

where the inequalities are clear when one compares (3.15) to (3.6) and (3.16) to (3.7). For example, if the true value of $\beta_2$ is 0, but $\theta_j^{**}$ is used as the estimate of $\theta_j$, $\hat{\beta}_2^{**}$ would be positive in a large sample. Thus, using the median PV produces estimates that overestimate the return to skills and underestimate the wage gap.

Though the PV methodology works when used properly to solve the measurement error problem, in practice it can be difficult to implement. In tests where plausible values are produced, few social scientists have the statistical training to understand the statistics behind the PV machinery, thus the reason many use either the 1st PV or the median PV. Even for social scientists who do understand the complicated estimation procedure required with PVs, the PV methodology only works when there is a primary analyst who produces
3.1. CURRENT METHODS

the plausible values. Very few tests (i.e., the NALS, NAEP, and TIMMS) have data sets which include such values. Producing these values can take quite a bit of time and so most data sets that include assessments only produce one estimate (generally the MLE) for each individual, making the PV methodology impossible to use in such cases.

There are also cases in which the data set may include plausible values, but those particular plausible values will not produce consistent estimates for the problem at hand. For example, Neal and Johnson (1996) use data from the National Longitudinal Study of Youth (NLSY) which includes a measure of cognitive skill taken from the Armed Forces Qualifying Test (AFQT). In the NLSY, the AFQT test was given to respondents at age 16-18 and then those respondents are followed. The wage data that Neal and Johnson (1996) use is measured at a later date, closer to when most of the individuals in the study are in their late 20s. The NLSY does not include plausible values for the AFQT, but for the purposes if this example, imagine it did. The PVs that would be included in the NLSY would be based on a conditioning model that included variables from when the individuals were 16-18 years of age. Thus, that conditioning model would not include certain variables like marital status at the current age. If marital status were included in the regression equation, then the AFQT PVs would need to be recalculated using marital status in the conditioning model. Thus, the PV methodology is not the most efficient for split sample designs or for longitudinal studies. As more information is gathered, the PVs will need to be re-estimated, something which is not done in the literature.

Marginal Maximum Likelihood Procedures

As noted in Chapter 2, one other procedure for implementing the marginal estimation model is the marginal maximum likelihood (MML) methods. MML methods are computationally more efficient than the PV methodology and presumably MML methods should solve the problem inherent in using PV methodology in longitudinal studies or split-sample designs.

As of this writing, the MML model has been implemented in software such as the AM software. The AM software, however, is not equipped to handle models in which \( \theta_j \) is an independent variable. The MML model (as used by the AM software) can only solve problems where the test score is the dependent variable. As is the case with the PV
methodology, a typical social scientist will not have the statistical know-how to produce results using the MML procedures without some sort of software.

3.2 Three Simulation Studies

Up until now, I have focused attention on the four current methods used to handle the measurement error inherent in the test score when the test score is used as an independent variable in a regression analysis. I have argued that ignoring the measurement error will cause bias to the regression coefficients. In particular, the coefficient on the observed test score will be biased toward zero and the coefficient on race will be biased away from zero. The EIV model does not ignore the measurement error, but is both unidentifiable and models the magnitude of the standard error as equal for every respondent in the study. I have also shown that the PV methodology produces consistent estimates, though some uses of the plausible values (most specifically the 1st plausible value and the median plausible value) will be problematic. Finally, I have noted that the MML model is not able to be implemented in any known software when $\theta_j$ is the independent variable in an analysis.

For the remainder of this chapter, I demonstrate through the use of simulation studies how the problems in the current models above manifest themselves. I remain focused on the issue of black-white wage gaps and use a rather simple regression model to examine the issue of the wage gap. In particular, I regress the log of wages on a measure of $\theta_j$ and an indicator variable for race.

3.2.1 Simulations Using Elementary Linear Regression

The simulations in this section are intended to show the extent of the bias in the regression coefficients when no measurement error is modeled. In all of the simulations, the true relationship of $w_j$, $\theta_j$, and $B_j$ is

$$TRUTH: w_j = 6.0 + 0.2\theta_j - 0.1B_j + \varepsilon_j$$

$$\varepsilon_j \sim N(0, \sigma^2),$$

(3.17)
3.2. THREE SIMULATION STUDIES

where \( w_j \) is the log of weekly wages, \( \theta_j \) is the true ability, \( B_j \) is an indicator variable for race that is 1 when the person is black and 0 when the person is white, and \( \varepsilon_j \) is random error. In my “truth”, blacks are paid about 10% less than whites even when controlling for ability. In addition, an increase in ability means an increase in wages and this relationship is linear.

Let \( N = 280 \). I set \( B_j \) such that the first 140 respondents are white and the second 140 respondents are black. Let \( \sigma^2 = 0.25 \) in all the simulations as well, meaning that \( \varepsilon \) is randomly generated from a \( N(0, 0.25) \) distribution. I randomly generate \( \theta_j \) and calculate \( w_j \) from the truth above. Since \( \theta_j \) is randomly generated from different distributions depending on the simulation I describe the generating distribution of \( \theta_j \) in more detail below.

Each simulation is analyzed 1000 times. The estimates shown in the tables below are the means of the 1000 estimates for each simulation. In all cases, the means, medians, and modes are very close to one another as the distribution of the 1000 estimates for each simulation is quite close to a normal distribution.

I randomly generate an error term for \( \theta_j \) that I call \( \nu_j \) which is generated from a \( N(0, 0.675) \). I then calculate \( \phi_j \) by summing \( \theta_j \) and \( \nu_j \). In the following three simulations, I estimate the regression equation \( w_j = \beta_0 + \beta_1 \phi_j + \beta_2 B_j + \varepsilon_j \).

In the first simulation (which I name Elementary Linear Regression-No Correlation or ELR-No Corr), \( \theta \) and race are uncorrelated. Thus, in ELR-No Corr, \( \theta_j \) is generated from a \( N(0, 1.25) \) distribution for both blacks and whites. (In the table, I specify the mean distribution of \( \theta \) for blacks and whites as \( \mu_{\theta,B} \) and \( \mu_{\theta,W} \) respectively. The standard deviation of the distribution for blacks and whites is \( \tau_{\theta,B} \) and \( \tau_{\theta,W} \) respectively.) I chose this distribution, because it centers \( \theta \) at 0 (which tends to be standard) and tends to draw \( \theta_j \)'s from a range of approximately \((-4, 4)\). In the second simulation, (which I name Elementary Linear Regression-Correlation 1 or ELR-Corr 1), I add correlation between race and \( \theta \) and set it to be \(-0.37\). Here, I generate \( \theta_j \) from a \( N(-0.5, 1.25) \) distribution for blacks and a \( N(0.5, 1.25) \) distribution for whites. In the third simulation, (which I name Elementary Linear Regression-Correlation 2 or ELR-Corr 2), I make the correlation between race and \( \theta \) stronger and set it to be \(-0.675\). I then generate \( \theta_j \) from a \( N(-1, 1.25) \) distribution for blacks and a \( N(1, 1.25) \) distribution for whites. In the first simulation, the estimation
Table 3.1: Results from Simulations-Bias in Elementary Linear Regression Coefficients when Measurement Error is Not Modeled

<table>
<thead>
<tr>
<th></th>
<th>( \beta_0 )</th>
<th>( \theta )</th>
<th>( B )</th>
<th>( \mu_{\theta,B} )</th>
<th>( \tau_{\theta,B} )</th>
<th>( \mu_{\theta,W} )</th>
<th>( \tau_{\theta,W} )</th>
<th>( \sigma^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Truth</strong></td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>0</td>
<td>1.25</td>
<td>0</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>ELR-No Corr</strong></td>
<td>Estimates</td>
<td>6.00</td>
<td>0.16</td>
<td>-0.10</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>0.28</td>
</tr>
<tr>
<td>( \rho = 0 )</td>
<td>S. E.</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Truth</strong></td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-0.50</td>
<td>1.25</td>
<td>0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>ELR-Corr1</strong></td>
<td>Estimates</td>
<td>6.01</td>
<td>0.16</td>
<td>-0.14</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>0.28</td>
</tr>
<tr>
<td>( \rho =-0.37 )</td>
<td>S. E.</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Truth</strong></td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-1.00</td>
<td>1.25</td>
<td>1.00</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>ELR-Corr2</strong></td>
<td>Estimates</td>
<td>6.05</td>
<td>0.16</td>
<td>-0.19</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>0.28</td>
</tr>
<tr>
<td>( \rho =-0.675 )</td>
<td>S. E.</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: \( N = 280 \) with 140 whites and 140 blacks. Each simulation run 1000 times. The truth is: \( \beta_0 = 6.0 \), \( \beta_1 = 0.20 \), \( \beta_2 = -0.10 \) for all three simulations. \( \phi_j = \theta_j + \nu_j \) where \( \nu_j \sim N(0,0.675) \). In ELR-No Corr, there is no correlation between \( \theta \) and \( B \), but in ELR-Corr1 and ELR-Corr2 there is. ELR-Corr1, the difference in the mean \( \theta \)s of the two groups is 1 whereas in ELR-Corr2, the difference in the mean \( \theta \)s of the two groups is 2.

results show what would be expected in an errors-in-variables regressions that has error on the \( \theta_j \) variable: that the estimate on the coefficient of \( \theta_j \) is biased toward 0. There is no effect on the estimates of the other coefficients, because they are not correlated with \( \theta \).

Table 3.1 shows that in the Elementary Linear Regression-No Correlation simulation, the estimate on the coefficient on \( \theta_j \) is 0.16 instead of 0.20. Neither of the other coefficients in the regression are affected by the measurement error.

In the second simulation, \( \theta \) and \( B \) are correlated such that on average, whites have higher cognitive skills than blacks. In this case, the estimation results show that not only is the estimate of the coefficient on \( \theta_j \) biased toward 0, but the estimate of the coefficient on race is biased away from 0. Table 3.1 shows that in the Elementary Linear Regression-Correlation 1 simulation, the estimate on the coefficient on \( \theta_j \) is again 0.16 instead of 0.20, though the coefficient on \( B_j \) is \(-0.14\) instead of \(-0.10\). Thus, when there is a correlation of \(-0.37\) between the \( \theta_j \) and race, the coefficient on race becomes more negative, suggesting a larger wage gap than actually exists.

The results are worse, the larger the difference in the mean of \( \theta \) for the two groups. In the third simulation, the estimation results show that the estimates are even more biased.
Table 3.1 shows that in the Elementary Linear Regression-Correlation 2 simulation, the estimate on the coefficient on $\theta_j$ is 0.16 instead of 0.20 and the coefficient on $B_j$ is $-0.19$ instead of $-0.10$. When the correlation between the race and $\theta_j$ is $-0.675$, the wage gap is estimated to be almost twice what it actually is.

This is clearly problematic! As noted in the introduction of this chapter, the elementary linear regression models is one of the most common models used to estimate racial wage gaps. However, when there is correlation between $\theta$ and race, the racial wage gap is overestimated by the elementary linear regression model. In the simulations, I set the truth such that blacks are paid 10% less than whites conditional on skills. However, the estimated coefficient on race in the elementary linear regression model implies that blacks are paid 19% less than whites conditional on skills. This could lead to inferences that policy work should focus more on reducing labor force discrimination rather than focusing some attention on reducing the racial skills gap.

### 3.2.2 Simulations Using Errors-in-Variables Model

A set of simulations is also likely to be useful in demonstrating the effects of using classic test theory to estimate $\theta_j$ instead of item response theory.

As in the simulations using the elementary linear model, the truth is (3.17). Let $N = 280$ with 140 blacks and 140 whites. Let $\sigma^2$ be 0.25 such that $\varepsilon$ is randomly generated from a $N(0, 0.25)$ distribution. In both of the simulations in this section, I generate $\theta_j$ from a $N(-1, 1.25)$ distribution for blacks and a $N(1, 1.25)$ distribution for whites. As before, each simulation is analyzed 1000 times and the estimates in the table are the means of the 1000 estimates for each simulation.

For the EIV model, the observed test score $\phi$ is the sum of the total correct on a test. In order to produce $\phi$, I generate a test of 10 items. I randomly generate the difficulty parameters of this test from a $N(1, 0.5)$ distribution and the discrimination parameters from a $Unif(0, 2)$ distribution. Note that the difficulty parameters are chosen such that if an IRT model were used, the standard errors on the estimated $\theta_j$s would be smaller for whites than for blacks. Using the 2-PL model, I randomly generate an item response for each examinee for each of the ten items where the item response was drawn from a
CHAPTER 3. CURRENT METHODS

Ber\(P(x_{ij} = 1|\theta_j, a_i, b_i)\) distribution where \(P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1+\exp(a_i(\theta_j-b_i))}\). Each examinee’s observed score \(\phi_j\) is calculated by summing the total correct for each examinee. I then assume that \(\theta_j\) is equal to a linear transformation of \(\phi_j\) plus some random error that I call \(\nu_j\).

I estimate the following EIV regression model in each of the EIV simulations:

\[
\begin{align*}
    w_j &\sim N(\beta_0 + \beta_1 \theta_j + \beta_2 B_j, \sigma_{\varepsilon\varepsilon}) \\
    \phi_j &\sim N(\delta + \Gamma \theta_j, \sigma_{\nu\nu}) \\
    \text{Cor}(B, \theta) &\neq 0 \\
    E(\epsilon|\phi, B) &= 0 \\
    \text{Cov}(\nu, \epsilon) &= 0.
\end{align*}
\]

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). The priors on the parameters in the regression equation are

\[
\begin{align*}
    \sigma_{\varepsilon\varepsilon} &\sim Unif(0, 1000) \\
    \beta_0 &\sim N(0, 10000) \\
    \beta_1 &\sim N(0, 10000) \\
    \beta_2 &\sim N(0, 10000).
\end{align*}
\]

The priors and hyperpriors on the parameters on \(\theta_j\) are

\[
\begin{align*}
    \theta_j|B_j = 1 &\sim N(\mu_{\theta,B}, \frac{1}{\tau_{\theta,B}}) \\
    \theta_j|B_j = 0 &\sim N(\mu_{\theta,W}, \frac{1}{\tau_{\theta,W}}) \\
    \mu_{\theta B} &\sim N(0, 1) \\
    \mu_{\theta W} &\sim N(0, 1) \\
    \tau_{\theta B} &\sim \Gamma(1, 1) \\
    \tau_{\theta W} &\sim \Gamma(1, 1).
\end{align*}
\]
A few notes on the priors and hyperpriors on $\theta_j$. First, note that the priors on $\theta_j$ are conditioned on race following Mislevy (1991). In addition, I set the hyperpriors on $\mu_{\theta,B}$, $\mu_{\theta,W}$, $\tau_{\theta,B}$, and $\tau_{\theta,W}$ such that the prior does not contain any information about whether or not blacks or whites have higher $\theta_j$.

The priors on $\Gamma$, $\delta$, and $\nu_j$ are different for the simulations. In the first simulation, I calculate $\Gamma$ and $\delta$ by estimating a simple linear regression equation, $\phi_j = \delta + \Gamma \theta_j + \nu_j$, using the true $\theta_j$. I calculate that, $\delta = 3.27$, $\Gamma = 1.45$ and $\sigma_{\nu,\nu}^2 = 1.45$.

Thus for the first simulation, I set $\delta = 3.27$, $\Gamma = 1.45$ and set a prior on $\sigma_{\nu,\nu}^2 \sim Unif(0,1000)$. I do this in order to get a “best-case scenario” set of results. I show these results in specification EIV-best in Table 3.2. For the second simulation, I set priors on $\delta$ and $\Gamma$ as well as on $\sigma_{\nu,\nu}^2$ as would be typical in an errors-in-variables analysis when $\theta_j$ is not known. In this simulation, the priors are

$$
\delta \sim N(3.27, 0.01)
$$

$$
\Gamma \sim N(1.45, 0.01)
$$

$$
\sigma_{\nu,\nu}^2 \sim Unif(0,1000).
$$

I set these priors such that the mean of the prior is the true value of $\delta$ and $\Gamma$, but added some variability (though the variability is quite small.) I show the results in specification EIV-prior in Table 3.2.

In both EIV models the mean $\theta$ of blacks is underestimated and the mean $\theta$ for whites is overestimated. In addition, the EIV models underestimate the standard deviation of $\theta$ for blacks, but overestimate the standard deviation of $\theta$ for whites.

These estimates on $\theta_j$ make some sense. Consider the test from which the observed test score is calculated. Recall that the 10 item test has difficulty parameters that are drawn from a $N(1,0.5)$ distribution, meaning that on average the 10 difficulty parameters fell between $(0,2)$. I use the 2-PL IRT model to calculate the probability with which an individual answers an item correctly. Recall from Chapter 2 that the difficulty parameter is equal to the value of $\theta_j$ when the probability of a correct response is 0.5.
### Table 3.2: Results from Simulations-Typical Errors-in-Variables Models

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\theta$</th>
<th>$B$</th>
<th>$\mu_{\theta,B}$</th>
<th>$\tau_{\theta,B}$</th>
<th>$\mu_{\theta,W}$</th>
<th>$\tau_{\theta,W}$</th>
<th>$\sigma^2_{\varepsilon,\varepsilon}$</th>
<th>$\delta$</th>
<th>$\Gamma$</th>
<th>$\sigma^2_{\nu,\nu}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Truth</strong></td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.25</td>
<td>3.27</td>
<td>1.45</td>
<td>1.4</td>
</tr>
<tr>
<td><strong>EIV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EIV Estimates</td>
<td>6.00</td>
<td>(0.05)</td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.14)</td>
<td>(0.18)</td>
<td>(0.04)</td>
<td>(0.26)</td>
<td>(0.21)</td>
<td></td>
</tr>
<tr>
<td>S. E.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC Error</td>
<td>0.07</td>
<td>(0.06)</td>
<td>(0.10)</td>
<td>(0.09)</td>
<td>(0.15)</td>
<td>(0.19)</td>
<td>(0.05)</td>
<td></td>
<td>(0.27)</td>
<td>(0.27)</td>
<td></td>
</tr>
<tr>
<td>EIV Prior</td>
<td>6.06</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.15)</td>
<td>(0.14)</td>
<td>(0.02)</td>
<td>(0.10)</td>
<td>(0.07)</td>
<td>(0.35)</td>
</tr>
<tr>
<td>S. E.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC Error</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.09)</td>
<td>(0.08)</td>
<td>(0.13)</td>
<td>(0.14)</td>
<td>(0.03)</td>
<td></td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.33)</td>
</tr>
</tbody>
</table>

Notes: $N = 280$ with 140 whites and 140 blacks. Each simulation run 1000 times. The truth is: $\beta_0 = 6.0$, $\beta_1 = 0.20$, $\beta_2 = -0.10$. In the EIV-best specification, the true $\Gamma$ and $\delta$ are used. In the EIV-Prior specification, priors are placed on $\Gamma$ and $\delta$. 

**CHAPTER 3. CURRENT METHODS**
Thus, for most of the blacks in the sample (whose $\theta_j$'s are randomly drawn from a $N(-1, 1)$ distribution), the probability of a correct response on any of the items is less than 50%. Thus, I expect the observed scores of blacks to have a smaller variability than the true variability of $\theta$ for blacks.

For whites, however, the 10 items are generated such that the variability of the observed score is far greater than the true score. Because the test is set up such that the difficulty parameter is so concentrated around 1 (which is the true mean of $\theta$ for whites), whites with very similar true $\theta_j$'s might have quite different observed scores. The EIV model assumes that the standard errors are similar in magnitude regardless of the difficulty of the test. Thus, the EIV model does not use the information about the difficulty level in the item parameters, producing an estimate of the standard deviation for whites that is far greater than the true standard deviation.

In EIV-best, the estimated regression coefficients replicate the truth exactly. This is to be expected as I use the true $\Gamma$ and $\delta$ in this model. However, in EIV-prior where I place priors on $\Gamma$ and $\delta$, the EIV model does very poorly at estimating the regression coefficients. The return to skills is underestimated and the black-white wage gap is estimated to be twice what it actually is, even though the priors use the true value of $\Gamma$ and $\delta$ as the mean. In EIV-prior, the estimates for $\Gamma$ and $\delta$ do not reflect the true values. Note as well, that in EIV-prior, the estimate of $\sigma_{\nu}^2$ is much lower than what it should be. It is likely that the EIV model has too much measurement error in its estimates of $\theta_j$ to produce good estimates of the regression coefficients. I discuss the role of measurement error in estimating the regression coefficients in more detail in Chapter 4.

Finally, in addition to the two simulation studies described above, I also calculated the Klepper-Leamer bounds (Klepper and Leamer, 1984) for the simulated data. I estimate bounds on the coefficient on race to be $(-0.2484, 0.2839)$. Clearly, as is the case with Bollinger’s (2003) estimates, the bounds are so large as to not be particularly useful.

### 3.2.3 Simulations Using Plausible Values Methodology

In Section 3.1.3, I described the method for obtaining regression coefficient estimates when using the PV methodology. In addition, I noted that problems with the estimates will arise
when a researcher uses only the first plausible value or the median plausible value as a fixed, independent variable. In the following three simulations, I produce a set of plausible values using the marginal estimation procedures. I then use those values in a simulation study to demonstrate that the PV methodology when used correctly, does produce consistent regression coefficients. A second simulation study shows that using only the first plausible value underestimates the standard errors. Finally, a third simulation study demonstrates that using the median plausible value biases the regression coefficients greatly.

As in the previous simulation studies the truth is (3.17). Let $N = 280$ with 140 respondents who are white and 140 respondents who are black. I let $\sigma^2$ be 0.25 and randomly generate $\varepsilon_j$ from a $N(0, 0.25)$ distribution. I randomly generate $\theta_j$ from a $N(-0.5, 1.25)$ distribution for blacks and a $N(0.5, 1.25)$ distribution for whites and calculate $w_j$ from the truth above. Each simulation is analyzed 1000 times and the estimates produced in the table are the means of the 1000 estimates for each simulation.

In order to produce a set of plausible values, I generate item responses $X_{ij}$ for each individual for a 10-item test using the true $\theta_j$s and a 2-PL model with discrimination parameters $a_i$ drawn from a $Unif(0, 2)$ and difficulty parameters $b_i$ generated from a $N(0, 0.8)$. Using the generated test data, I draw five plausible values from the posterior distribution $p(\theta_j | X_{ij}, B_j, a_i, b_i)$ where

$$p(\theta_j | X_{ij}, B_j, a_i, b_i) \propto p(X_{ij} | \theta_j, a_i, b_i)p(\theta_j | B_j).$$  

WinBUGS software is used to calculate the posterior distribution and then draw the five imputation. The priors on the regression coefficients are the same as in the EIV model (3.19). I set the generated item parameters as fixed and did not re-estimate them. (This seems a reasonable assumption given that in most surveys that use the PV methodology like the 1992 National Adult Literacy Survey, the item parameters are estimated based on thousands of responses and so the measurement error is often almost negligible on these parameters.) Unlike in the PV methodology used in large surveys, I use a small conditioning model, and only condition $\theta_j$ on race, as that is the only covariate in the regression equation. The priors on $\theta_j$ are the same as in the EIV model, (3.20).
I estimate three separate sets of regression coefficients in each of the simulations. In the first simulation (called Plausible Value or PV), I estimate the regression coefficients using the recommended PV procedures described in Section 3.1.3. I estimate the regression equation \( w_j = \beta_0 + \beta_1 PV(q) + \beta_2 B_j + \varepsilon_j \) five times each time using a different plausible value \( PV(q) \) where \( q = 1, \ldots, 5 \). I estimate the standard errors on the coefficients through (3.13).

In the second simulation (called 1st Plausible Value or 1st PV), I estimate the regression equation \( w_j = \beta_0 + \beta_1 PV_1 + \beta_2 B_j + \varepsilon_j \) using the first plausible value as my estimate for \( \theta_j \). Standard errors for the estimates of the regression coefficients only include the model based uncertainty and no Monte Carlo uncertainty is accounted for.

Finally, in the third simulation (called Median Plausible Value or Med PV) I estimate the regression equation \( w_j = \beta_0 + \beta_1 med(PV) + \beta_2 B_j + \varepsilon_j \) where \( med(PV) \) is the median plausible value. Again, standard errors only include the model based uncertainty. The results of the three sets of simulations are shown in Table 3.3.

As expected, the results in Table 3.3 show that the PV methodology (when used properly) works well to estimate the coefficients in the regression. In both the case of the PV and the 1st PV simulations, the point estimates of the regression coefficients are quite close to the truth. Note that in the 1st PV simulation the standard errors are smaller than in...
PV, because 1st PV does not take into account the Monte Carlo error that results from drawing the plausible values from the posterior distribution. The PV simulation does take the Monte Carlo error into account (at least up to five draws worth of error).

Using the median plausible value is much more problematic as can be seen in Med PV in Table 3.3. As expected, using the median plausible value as the estimate of $\theta_j$ shrinks the $\theta$ distribution toward its group means. Note that the estimated standard deviations of the $\theta$ distributions are underestimated. This further biases the regression coefficient results. In particular, the estimate on race is underestimated suggesting a smaller race gap than actually exists. In addition, the return to skills is overestimated.

Thus, while the plausible value methodology can work to estimate consistent regression coefficients, the estimates must be calculated using (3.12) and the standard errors must be calculated using (3.13). In addition, the conditioning model must be sure to include all the covariates in the regression equation. In cases where the data for the other independent variables are collected after the PVs are calculated, the PVs will need to be recalculated. Without these procedures, standard errors could be underestimated or worse, bias could result in the regression coefficients.

3.3 Conclusion

Four methods for handling the measurement error problem were described in this chapter: ignoring the measurement error (classic ELR), classic errors-in-variables (EIV), the Plausible Value (PV) Methodology, and marginal maximum likelihood (MML) methods. Simulations showed the large bias that can result from ignoring the measurement error completely in ELR. While classic EIV methods will model the measurement error, EIV models are typically unidentifiable and are unable to model an error structure that is different for each individual. The PV methodology will accurately model the error when the conditioning model includes all the covariates in the regression equation. However, the PV methodology can only be used by the typical social scientist when a primary analyst produces the plausible values. In addition, for split sample and longitudinal data designs, the PVs will have to be re-estimated every time new data is collected in order to produce
consistent estimates. The only software which implements the MML model does not allow for $\theta_j$ to be an independent variable.

In the next chapter I introduce a new model, the Mixed Effects Structural Equations (MESE) Model which models $\theta_j$ as an independent variable. The MESE model includes measurement error on $\theta_j$ in order to correct for the bias when the measurement error is ignored. In addition, the MESE model utilizes IRT in order to model the error as different for each individual unlike the classical EIV models which model the magnitude of the error as the same for each individual. Using the IRT model also helps to handle some of the identifiability issues of the typical EIV model. In addition, the MESE model does not require a set of plausible values to be produced, making it more universal than the PV methodology.
Chapter 4

The Mixed Effects Structural Equations Model: A New Method

In Chapter 3, I introduce a typical regression equation used in examining black-white wage gaps,

\[ w_j = \beta_0 + \beta_1 \theta_j + \beta_2 B_j + \varepsilon_j \]
\[ \varepsilon_j \sim N(0, \sigma^2), \]

where \( w_j \) is log wages, \( \theta_j \) is a latent variable defined as ability, and \( B_j \) is a dummy variable that is 1 when the individual is black and 0 when the individual is white. I noted that all four of the current methods for handling the measurement error inherent in \( \theta_j \) have limitations. Ignoring the measurement error in Elementary Linear Regression (ELR) analysis produces biased regression coefficients. Typical Errors-In-Variables (EIV) analysis is not only unidentifiable, but also models homoskedastic measurement error. The Plausible Value (PV) methodology works, but only in cases where a primary analyst produces a set of plausible values and no new data is collected after the plausible values are produced. Marginal Maximum Likelihood (MML) methods have potential to work, but to date have only been used in cases where cognitive skills are the dependent variable.

I have developed the Mixed Effects Structural Equation (MESE) Model to address these issues. The MESE model places \( \theta_j \) as an independent variable in a regression, appropriately
models the measurement error in a heteroskedastic fashion, deals with the identifiability issues of the EIV model, and can be used for any set of assessments or tests without the need for calculation of plausible values.

In this chapter, I define the MESE model. I specify the model for the problem of examining the issue of black-white wage gaps. I investigate the issue of identifiability noting that the arbitrary nature of the scale of $\theta$ and the estimate of the regression coefficient on $\theta_j$ are unidentifiable, but that setting the scale of $\theta$ solves this problem. I examine a number of assumptions and properties of the MESE model noting where bias on the estimates of the regression coefficients occurs. I find bias can occur on the regression coefficients when the prior on $\theta_j$ is not conditioned on the other covariates in the regression equation. Additionally, when the measurement error on $\theta_j$ is high, the regression coefficients estimates are not consistent. I find small bias on the regression coefficients when the prior on $\theta$ is assumed to be symmetric but the true distribution of $\theta$ is skewed. I find very little bias on the regression coefficients when the IRT model is misspecified. I also provide recommendations for the length of the test and the variability of the items in order to avoid bias in the regression coefficients. Finally, I compare simulation results using the MESE model with the current methods described in Chapter 3. I note how much better the results from the MESE model are than using an ELR model or the EIV model. I show similar results for the MESE model and the PV methodology. I note that the MESE model is more universal than the PV methodology. The MESE model can be used for any data set that does not calculate plausible values as well as in data sets that have split survey designs or are longitudinal in nature without requiring the calculation of new PVs.

4.1 The Mixed Effects Structural Equations Model

The basic MESE model can be written in hierarchical form as

$$\theta_j | Y_j \sim p(\theta_j | Y_j, \tau_j^2)$$ (4.1)

$$X_{ij} | \theta_j \sim IRT(X_{ij} | \theta_j, \gamma_i)$$ (4.2)

$$w_j | Y_j, \theta_j \sim N(\beta_0 + \beta_1 \theta_j + \beta_2 Y_j, \sigma^2),$$ (4.3)
4.1. THE MIXED EFFECTS STRUCTURAL EQUATIONS MODEL

where \( X_{ij} \) is the item responses of individual \( j \) to item \( i \), \( Y_j \) are the set of background variables (which when examining black-white wage gaps includes \( B_j \), an indicator variable that is 1 if the respondent is black and 0 if the respondent is white), \( w_j \) is the dependent variable in regression, \( IRT(X_{ij}|\theta_j, \gamma_i) \) is the IRT model used to score the test and \( \tau_j^2 \) and \( \gamma_i \) are parameters in the conditioning model and the IRT model respectively.

The equation of primary interest is

\[
 w_j = \beta_0 + \beta_1 \theta_j + \beta_2 B_j + \varepsilon_j. 
\]

The estimates of skills (\( \theta_j \)) from the IRT model are themselves noisy, but they can be treated as random variables in a mixed-effects regression.

Together (4.1), (4.2), and (4.3) combine to give the following likelihood for the MESE model

\[
 L(w_1, w_2, \ldots, w_n | \theta_1, \theta_2, \ldots, \theta_n, Y_1, Y_2, \ldots, Y_n, x_{1,1}, x_{1,2}, \ldots, x_{m,n}) = \prod_j P(\theta_j | Y_j) \times \prod_j \prod_i P(x_{i,j} | \theta_j)^{x_{i,j}} (1 - P(x_{i,j} | \theta_j))^{1-x_{i,j}} \times \prod_j \Phi(\frac{w_j - (\beta_0 + \beta_1 \theta_j + \beta_2 Y_j)}{\sigma}). \]

The conditioning model, (4.1) often will be assumed to have a normal distribution. In practice, either priors or point-estimates are additionally needed for \( \tau^2 \), \( \sigma^2 \) and \( \gamma \). In addition, priors are needed for the \( \beta \)'s. In most cases, these will be assumed to be normally distributed with mean 0. One way to estimate the MESE model is to use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (WinBUGS software is suggested as one way to implement the model in practice).

In the case of analyzing black-white wage gaps (and many other social science issues) the marginal posterior estimates of the \( \beta \)s will be of most interest as they are the regression coefficients. In order to estimate the \( \beta \)s, the latent variables \( \theta_j \) must be integrated out of the likelihood.

One advantage of the MESE model is its ability to be defined in numerous ways, both as a mixed effects model and as a structural equations model. Unlike typical ELR models where skills, \( \theta_j \), are modeled as a non-stochastic explanatory variable, in the MESE model
skills are modeled as a random variable. The advantages of modeling skills as a random effect are two-fold. First, by assuming skills are a random effect, skills are modeled as having error. Since Chapter 2 demonstrated that all test scores have error associated with them, this is an appropriate modeling choice. In addition, though, modeling skills as a random effect, also has the advantage of allowing the skills of the individuals in the study to be considered to be a random sample from a very large population. When using the MESE model to understand the effect of skills on certain outcome variables, this is important. Most researchers will be less interested in the effect skills has on the dependent variable for any particular individual in the study, but rather will want to make inferences across the entire infinite population of skill levels.

4.1.1 The MESE Model as a Structural Equations Model

The MESE model can also be defined as a structural equations model. Consider a more general form of structural equations models with latent variables presented by Bollen (2002):

\[
\eta = \alpha_\eta + B_\eta + \Gamma_\xi + \zeta \tag{4.5}
\]
\[
w = \alpha_w + \Lambda_\eta \eta + \varepsilon \tag{4.6}
\]
\[
X = \alpha_X + \Lambda_\xi \xi + \delta, \tag{4.7}
\]

where \(\eta\) is a vector of latent endogenous variables, \(\xi\) is a vector of latent exogenous variables, \(Y\) is a vector of observed variables associated with the latent endogenous variables, \(X\) is a vector of observed variables associated with the latent exogenous variables, and \(\zeta, \varepsilon, \delta\) are all vectors of errors with mean 0 that are uncorrelated with one another. Equation (4.5) is called the latent variable model, (4.6) is called the structural model and (4.7) is called the measurement model.

If I modify equations (4.5)–(4.7) as follows, I have the MESE model.

\[
\eta = \xi \tag{4.8}
\]
\[
w = \alpha_w + \Lambda_\eta \eta + \varepsilon \tag{4.9}
\]
\[
\log(1 - E(X)) - \log(E(X)) = \alpha_X + \Lambda_\xi \xi + \delta. \tag{4.10}
\]
I simplify the latent variable equation (4.5) as shown in equation (4.8), because there is only one latent variable: skills as measured by a cognitive test. The structural model (4.6) does not change. Equation (4.7) is modified to become equation (4.10) to reflect the Bernoulli generalized linear model (glm) for measuring the latent variable with the IRT model (here the IRT model is the 2-PL). In Fox and Glas (2003) they use the Multilevel IRT (MLIRT) model. Their model attempts to control for a latent variable (e.g., IQ) to predict how student performance on a test may be different for schools that have been given treatments and schools that have not. They use a normal ogive model instead of the 2-PL model to model the latent variable on the predictor side of the equation.

The advantage of having the MESE model be a structural equations model (SEM) is that SEMs are known for their ability to model constructs as latent variables. Since skills are a latent variable, modeling them in a SEM framework allows me to effectively capture the unreliability of the measurement of skills in the model, which allows the structural relationships between skills and the other variables in the model to be accurately estimated.

### 4.1.2 Identifiability of the MESE Model

The MESE model solves many of the problems of identifiability that the typical EIV model has. As noted in Section 3.1.2, the typical errors-in-variables model is unidentifiable unless there is additional data in one of three areas: (1) replicate measure of $\phi_j$, (2) distributional information on $\theta$, or (3) distributional or moment restrictions on the error distribution (Dynarski, 1999). Fortunately, the MESE model actually gives some information in all three areas. I examine each one individually.

1. Replicate measures of $\phi_j$.

While, the EIV model uses one total test score as its measure of $\phi_j$, the MESE model can be thought to have $M$ measures of $\phi_j$ where $M$ is the number of items on the test. So for a test that is longer than 1 item, replicate measures of $\phi_j$ are available in the MESE model, even if those replicate measures are rather crude. Individually each item response is a poor measure of $\theta_j$; together they are stronger. Using IRT instead of CTT allows the MESE model to exploit the fact that a test is a collection of items, rather than one large measurement device.
(2) Distribution data on \( \theta \).

The MESE model, through the IRT model produces a posterior population estimates of \( \theta \) given the item responses. This posterior distribution is exploited to provide distributional data on \( \theta \). Posterior means and standard deviations provide estimates of the mean and standard deviation of the distribution of \( \theta \).

(3) Distributional data on the errors.

The MESE model does not require any of the six assumptions listed in Section 3.1.2 about the measurement error and/or the regression error that will make the EIV model identifiable. The information function of the IRT model provides data about the measurement error. Section 2.5 described the inverse relationship between the information function and the standard error of the estimate of \( \theta_j \) where \( SE(\hat{\theta}_j) = \frac{1}{I(\theta_j)} \). This function is exploited in the MESE model to provide information about the standard error of \( \theta_j \).

Thus the MESE model solves the three major identifiability issues in the EIV model. However, the well-known identifiability issues in the IRT model are present in the MESE model as well. The location and scale parameters of the latent distribution are confounded with the difficulty and discrimination parameters of the measurement model. In the 2-PL model, the scale of the difficulty parameter and the scale of \( \theta_j \) are identified only up to an additive constant. In addition, the scale of the discrimination parameter and the scales of both \( \theta_j \) and the difficulty parameter are identified only up to a multiplicative constant. The following example will demonstrate this.

Let there be a 1-item test taken by individual \( j \) and let \( \theta_j^{***} = \zeta(\theta_j + \rho) \), let \( b^{***} = \zeta(b + \rho) \), and let \( a^{***} = \frac{a}{\zeta} \) where \( a \) is the discrimination parameter in the IRT model, \( b \) is the difficulty parameter in the IRT model, \( \zeta \) is a multiplicative constant, and \( \rho \) is an additive constant. Then

\[
P(x_j|\theta_j, a, b) = \frac{1}{1 + \exp(-1.7a(\theta_j - b))} = \frac{1}{1 + \exp(-1.7a^{***}(\theta_j^{***} - b^{***}))}.
\]

This indeterminacy of the IRT scales has further ramifications for the regression coefficients.
4.1. THE MIXED EFFECTS STRUCTURAL EQUATIONS MODEL

For example, in a regression with no intercept, let

\[ w_j = \beta_1 \theta_j + \beta_2 B_j + \varepsilon_j, \quad (4.12) \]

where \( w_j \) is the log weekly wage of person \( j \), \( \theta_j \) is the true ability of person \( j \) and \( B_j \) is an indicator variable that is equal to 1 when person \( j \) is black and equal to 0 when person \( j \) is white. Let \( \bar{w}_B \) and \( \bar{w}_W \) be the mean of log wage for blacks and whites respectively, and \( \bar{\theta}_B \) and \( \bar{\theta}_W \) be the corresponding means for the skills measure. Define the estimated covariances of \( \theta \) and \( w \) for the two demographic groups to be \( \hat{\sigma}_{\theta w, B} \) and \( \hat{\sigma}_{\theta w, W} \), and the estimated variance of \( \theta \) for the groups to be \( \hat{\sigma}_{\theta \theta, B} \) and \( \hat{\sigma}_{\theta \theta, W} \). Finally let \( \hat{\eta} \) be the fraction of the sample that is black. In Chapter 3, I showed that (3.4) and (3.5) are the OLS estimates in a standard linear regression model.

\[ \hat{\beta}_1 = \frac{\hat{\eta} \hat{\sigma}_{\theta w, B} + (1 - \hat{\eta}) \hat{\sigma}_{\theta w, W}}{\hat{\eta} \hat{\sigma}_{\theta \theta, B} + (1 - \hat{\eta}) \hat{\sigma}_{\theta \theta, W}} \]

and

\[ \hat{\beta}_2 = (\bar{w}_B - \bar{w}_W) - \hat{\beta}_1 (\bar{\theta}_B - \bar{\theta}_W). \]

Now let \( \theta_j^{***} = \alpha (\theta_j + \rho) \). Then, when I estimate the coefficients in (4.12),

\[ \hat{\beta}_1^{***} = \frac{\hat{\eta} \alpha \hat{\sigma}_{\theta^{***} \theta^{***}, B} + (1 - \hat{\eta}) \alpha \hat{\sigma}_{\theta^{***} \theta^{***}, W}}{\hat{\eta} \alpha^2 \hat{\sigma}_{\theta^{***} \theta^{***}, B} + (1 - \hat{\eta}) \alpha^2 \hat{\sigma}_{\theta^{***} \theta^{***}, W}}, \]

making \( \hat{\beta}_1^{***} = \frac{\hat{\beta}_1}{\alpha} \).

The estimated coefficient for \( \beta_2 \) should not change because:

\[ \hat{\beta}_2^{***} = (\bar{w}_B - \bar{w}_W) - \frac{\hat{\beta}_1}{\alpha} \alpha (\bar{\theta}_B - \bar{\theta}_W) = \hat{\beta}_2. \]

Thus the indeterminacy of IRT scales will have an effect on the estimate of the regression coefficient on \( \theta_j \). No effect will be had on the estimated coefficient on race. When the researcher is only interested in measuring the black-white wage gap and is not interested in measuring the return to skills, the indeterminacy of the IRT scales poses no issues.
The indeterminacy of the IRT scales can be avoided. One common way is to assume that the mean of the latent distribution of $\theta$ is 0 and the variance is 1, thus fixing the scale to be in terms of standard deviation units. This is often done on most tests during the scaling phase of the test. Test scores are often reported in a “reporting scale” rather than in the “theta scale.” In the case of the MESE model, it may be useful to use the “theta scale” in order to have a standardized understanding of what the $\beta_1$ estimate means.

Throughout this discussion of identifiability, I have been assuming that the item parameters (discrimination, difficulty, and guessing) are known and fixed. This is often a reasonable assumption as the item parameters are often estimated and published with the test. In some circumstances, the IRT parameters are estimated using more than those individuals who participated in the that particular version of the test. For example, in the case of the 1992 National Adult Literacy Survey (NALS), many of the items on the test were taken from other assessments including the 1985 Young Adult Literacy Assessment (YALA) and the 1990 Workplace Literacy Survey (WLS). Thus, the reported item parameters were estimated not just with those who participated in the 1992 NALS, but also the respondents of the 1985 YALA, and the 1990 WLS. The published item parameters of the 1992 NALS are likely be more accurately estimated than any estimates a secondary analyst could determine on her own. However, there may be cases in which the item parameters would need to be estimated in the MESE model as well. In such cases, one of the three methods for estimating IRT models described in Chapter 2 will be useful, though given the Bayesian nature of the MESE model, Patz and Junker’s (1999) MCMC methods are recommended.

4.2 The Estimates Produced by the MESE Model

Up until now, I have described the MESE model as a hierarchical model of the form

\[
\theta_j | Y_j \sim p(\theta_j | Y_j, \tau_j^2)
\]

\[
X_{ij} | \theta_j \sim IRT(X_{ij} | \theta_j, \gamma_i)
\]

\[
w_j | Y_j, \theta_j \sim N(\beta_0 + \beta_1 \theta_j + \beta_2 Y_j, \sigma^2),
\]
categorized it as a structural equations model and noted that the model is unidentifiable without fixing the scale of $\theta$. I will now turn my attention to examining how well the MESE model performs in certain situations. Before I examine any particular situation, I will calculate $\hat{\beta}_2$ that the MESE model produces. With a few simplifying assumptions, I can produce an estimate of the mean and variance of the marginal posterior distribution of the coefficient on race $\beta_2$. With these estimates, I am better able to examine situations where bias may occur.

Following directly from the likelihood (4.4), the joint posterior distribution of $\beta$ is

$$p(\beta_0, \beta_1, \beta_2| w_1, \ldots, w_n, Y_1, \ldots, Y_n, x_{1,1}, \ldots, x_{m,n}) \propto \prod_j P(\theta_j|Y_j) \Phi \left( \frac{w_j - (\beta_0 + \beta_1 \theta_j + \beta_2 Y_j)}{\sigma} \right) \prod_j \prod_i P(x_{ij}|\theta_j)^{x_{ij}} (1 - P(x_{ij}|\theta_j))^{1-x_{ij}} \pi(\beta),$$

where $\pi(\beta)$ is the prior placed on the vector of regression coefficients $\beta$. The constant of proportionality is found by integrating out $\beta_0$, $\beta_1$, $\beta_2$, and all $\theta_j$s.

Unfortunately, even assuming that $P(\theta_j|Y_j)$ and the priors on $\beta$ are normal, when $P(x_{ij}|\theta_j)$ is any of the logistic IRT models, there is no closed form solution for the joint posterior of $\beta_0, \beta_1, \beta_2$. For example, let

$$\pi(\beta_k) \sim N(0, \xi_k^2)$$

$$P(\theta_j|Y_j) \sim N(\mu_Y, \tau_Y^2).$$

Then in the case of one student and one test item scored using the 2-PL IRT model, the joint posterior for the $\beta$s looks like

$$f(\beta_0, \beta_1, \beta_2| w, X, Y) \propto$$

$$\int_\theta exp \left( -\frac{1}{2} \left( \frac{(w - (\beta_0 + \beta_1 \theta + \beta_2 Y))^2}{\sigma^2} + \frac{(\theta - \mu_Y)^2}{\tau_Y^2} + \frac{\beta_0^2}{\xi_0^2} + \frac{\beta_1^2}{\xi_1^2} + \frac{\beta_2^2}{\xi_2^2} + \frac{\beta_0 \beta_1}{\xi_0 \xi_1} + \frac{\beta_0 \beta_2}{\xi_0 \xi_2} + \frac{\beta_1 \beta_2}{\xi_1 \xi_2} \right) \right) +$$

$$\left( 1 (X = 1) \frac{1}{1 + exp(a(\theta - b))} + 1 (X = 0) \left( 1 - \frac{1}{1 + exp(a(\theta - b))} \right) \right) \, d\theta.$$

This is clearly not proportional to any standard exponential family model as a function of the $\beta$s and the situation will not improve with more test questions or persons (except in
terms of asymptotics).

Fortunately, with a few simplifying assumptions, insight into the MESE model estimates (in particular the estimate of $\beta_2$) is possible. As I showed in Section 4.1.2, the estimate for $\beta_1$ is dependent on the scale of $\theta_j$ and so my first simplifying assumption will be to set $\beta_1 = 1$. I can make this assumption without loss of generality because $\theta_j$ is a latent variable and thus its scale is entirely arbitrary. Second, in order to further simplify, I replace the assumption that $X_j | \theta_j \sim \text{Bernoulli}(P(\theta_j))$, (4.14) with $X_j | \theta_j \sim N(\theta_j, \lambda^2)$. (4.15)

This seems a reasonable thing to do if $X_j$ is thought of as being the observed test score that is scored using the IRT model rather than a response vector. The third assumption is that $\beta_0 = 0$. This assumption is easy to implement by simply mean centering each variable in the regression. For notation purposes, I will call $\tilde{\theta}_j = \theta_j - \bar{\theta}$, $\tilde{X}_j = X_j - \bar{X}$, and $\tilde{w}_j = w_j - \bar{w}$. Making this assumption does result in the loss of the variability of estimating $\beta_0$, but for our purposes here, this will be acceptable.

Finally, I continue to look at one particular case of the MESE model, namely when $Y_j$ includes only one variable which is $B_j$, an indicator variable that is 0 when the person is white and is 1 when the person is black. I assume that the standard deviation of $\theta_j | B_j = \text{black}$ and $\theta_j | B_j = \text{white}$ is the same and call it $\tau^2$. Let the mean of $\theta_j | B_j = \text{black}$ be $\mu_{\theta,B}$ and the mean of $\theta_j | B_j = \text{white}$ be $\mu_{\theta,W}$.

Our simplified MESE model is then:

$$\tilde{w}_1, ..., \tilde{w}_N | \tilde{\theta}_1, ..., \tilde{\theta}_N, B_1, ..., B_N, \beta_2, \sigma^2 \sim \text{IND} N(\tilde{\theta}_j + \beta_2(B_j - \bar{B}), \sigma^2)$$ (4.16)

$$\tilde{\theta}_1, ..., \tilde{\theta}_N | B_1, ..., B_N, \mu_{\theta,B}, \mu_{\theta,W}, \tau^2 \sim \text{iid} N(\mu_{\theta,W} + (\mu_{\theta,B} - \mu_{\theta,W})B_j, \tau^2)$$

$$\tilde{X}_1, ..., \tilde{X}_N | \tilde{\theta}_1, ..., \tilde{\theta}_N, \tilde{\lambda}^2 \sim \text{IND} N(\tilde{\theta}_j, \lambda^2)$$

$$\beta_2 \sim N(\alpha, \xi_2^2).$$

With this simplified MESE model, I can determine analytically the mean and variance of
the posterior distribution of $\beta_2$, which I do below in Theorem 4.2.1. Once I have determined a mean and variance of the posterior distribution, I will examine how not conditioning the prior on $\theta_j$ on $B_j$ and increasing the measurement error will affect the estimate of $\beta_2$.

**Theorem 4.2.1.** When using the MESE model in (4.16), the posterior distribution of $\beta_2$ is normal distribution with mean

$$
\mu_{\beta_2} = \frac{\xi_2^2 \left( (\lambda^2 + \tau^2) \sum_j \tilde{w}_j (B_j - \overline{B}) - \tau^2 \sum_j \tilde{X}_j (B_j - \overline{B}) \right)}{(\xi_2^2 \lambda^2 + \xi_2^2 \tau^2) \sum_j (B_j - \overline{B})^2 + \tau^2 \lambda^2 + \sigma^2 \lambda^2 + \sigma^2 \tau^2},
$$

and variance, $\sigma_{\beta_2}^2$:

$$
\sigma_{\beta_2}^2 = \frac{\xi_2^2 \left( \tau^2 \lambda^2 + \sigma^2 \lambda^2 + \sigma^2 \tau^2 \right)}{\tau^2 \lambda^2 + \sigma^2 \lambda^2 + \sigma^2 \tau^2 + (\xi_2^2 \lambda^2 + \xi_2^2 \tau^2) \sum_j (B_j - \overline{B})^2}.
$$

**Proof.** The likelihood of the simplified MESE model (4.16) is

$$
p(\tilde{w}_1, ..., \tilde{w}_N, \tilde{\theta}_1, ..., \tilde{\theta}_N, \tilde{X}_1, ..., \tilde{X}_N | B_1, ..., B_N, \beta_2, \sigma^2, \mu_{\theta,W}, \mu_{\theta,B}, \tau^2, \lambda^2) =
\prod_j \frac{1}{\sqrt{2\pi \sigma^2}} \times \exp\left( -\frac{1}{2\sigma^2} \left( \tilde{w}_j - (\tilde{\theta}_j + \beta_2 (B_j - \overline{B})) \right)^2 \right) \times
\frac{1}{\sqrt{2\pi \lambda^2}} \times \exp\left( -\frac{1}{2\lambda^2} \left( \tilde{X}_j - \tilde{\theta}_j \right)^2 \right) \times
\frac{1}{\sqrt{2\pi \tau^2}} \times \exp\left( -\frac{1}{2\tau^2} \left( \tilde{\theta}_j - (\mu_{\theta,W} + (\mu_{\theta,B} - \mu_{\theta,W}) B_j) \right)^2 \right).
$$

Since $\theta_j$ is a latent variable, I must integrate out $\theta_j$. In order to do so, I multiply out the terms of the likelihood, combine them, complete the square in $\tilde{\theta}_j$ and then isolate the terms that include $\tilde{\theta}_j$. 
The multiplied out likelihood with the $\tilde{\theta}_j$ terms grouped in the last two lines is

$$
p\left(\tilde{w}_1,...,\tilde{w}_N,\tilde{\theta}_1,...,\tilde{\theta}_N,\tilde{X}_1,...,\tilde{X}_N|B_1,...,B_N,\beta_2,\sigma^2,\mu_{\theta,B},\mu_{\theta,W},\tau^2,\lambda^2\right) = \prod_j \sqrt{\frac{\tau_j^2 \lambda^2 + \sigma^2 \lambda^2 + \sigma^2 \tau^2}{(2\pi)^2}} \times \exp \left( -\frac{1}{2} \left( \frac{\tilde{w}_j^2 - 2\beta_2 \tilde{w}_j (B_j - \bar{B}) + \beta_2^2 (B_j - \bar{B})^2}{\sigma^2} \right) \right)
$$

$$
\times \exp \left( -\frac{\tau_j^2 \lambda^2}{2\pi \sigma^2 \lambda^2} - \frac{2\mu_{\theta,W} \mu_{\theta,B} B_j - 2\mu_{\theta,B}^2 B_j^2 + 2\mu_{\theta,W} \mu_{\theta,B} B_j^2 + \mu_{\theta,W}^2 B_j^2 + \frac{\tilde{X}_j^2}{\lambda^2} - \frac{\left(\tilde{w}_j \tau_j^2 \lambda^2 - \beta_2 (B_j - \bar{B}) \tau_j^2 \lambda^2 + \mu_{\theta,W} \sigma^2 \lambda^2 + (\mu_{\theta,B} - \mu_{\theta,W}) B_j \sigma^2 \lambda^2 + \tilde{X}_j \sigma^2 \tau^2\right)^2}{\sigma^2 \tau_j^2 \lambda^2 (\tau_j^2 \lambda^2 + \sigma^2 \lambda^2 + \sigma^2 \tau^2)} \right) \right)
$$

I assume that $\alpha$, $\sigma^2$, $\mu_0$, $\mu_1$, $\tau^2$, $\lambda^2$, and $\xi_j^2$ are all fixed. Then

$$
p(\tilde{w}_1,...,\tilde{w}_N,\tilde{X}_1,...,\tilde{X}_N|B_1,...,B_N,\beta_2,\sigma^2,\mu_{\theta,B},\mu_{\theta,W},\tau^2,\lambda^2) = \int \cdots \int p(\tilde{w}_1,...,\tilde{w}_N,\tilde{\theta}_1,...,\tilde{\theta}_N,\tilde{X}_1,...,\tilde{X}_N|B_1,...,B_N,\beta_2,\sigma^2,\mu_{\theta,B},\mu_{\theta,W},\tau^2,\lambda^2) d\tilde{\theta}_1 \cdots d\tilde{\theta}_N.
$$

The last two lines of (4.19) are the only terms in the likelihood that involve $\theta_j$. Thus, when integrating in terms of $\theta_j$ the first three lines will fall out of the integral. These three latter terms of (4.19) then integrate to 1, because they are the pdf of a normal distribution for each $\tilde{\theta}_j$. The likelihood that does not involve $\theta_j$ is

$$
p\left(\tilde{w}_1,...,\tilde{w}_N,\tilde{X}_1,...,\tilde{X}_N|B_1,...,B_N,\beta_2,\sigma^2,\mu_{\theta,B},\mu_{\theta,W},\tau^2,\lambda^2\right) = \prod_j \sqrt{\frac{\tau_j^2 \lambda^2 + \sigma^2 \lambda^2 + \sigma^2 \tau^2}{(2\pi)^2}} \times \exp \left( -\frac{1}{2} \left( \frac{\tilde{w}_j^2 - 2\beta_2 \tilde{w}_j (B_j - \bar{B}) + \beta_2^2 (B_j - \bar{B})^2}{\sigma^2} \right) \right)
$$

$$
\times \exp \left( -\frac{\tau_j^2 \lambda^2}{2\pi \sigma^2 \lambda^2} - \frac{2\mu_{\theta,W} \mu_{\theta,B} B_j - 2\mu_{\theta,B}^2 B_j^2 + 2\mu_{\theta,W} \mu_{\theta,B} B_j^2 + \mu_{\theta,W}^2 B_j^2 + \frac{\tilde{X}_j^2}{\lambda^2} - \frac{\left(\tilde{w}_j \tau_j^2 \lambda^2 - \beta_2 (B_j - \bar{B}) \tau_j^2 \lambda^2 + \mu_{\theta,W} \sigma^2 \lambda^2 + (\mu_{\theta,B} - \mu_{\theta,W}) B_j \sigma^2 \lambda^2 + \tilde{X}_j \sigma^2 \tau^2\right)^2}{\sigma^2 \tau_j^2 \lambda^2 (\tau_j^2 \lambda^2 + \sigma^2 \lambda^2 + \sigma^2 \tau^2)} \right) \right).
Using Bayes rule,

\[
p(\beta_2 | \tilde{w}_1, ..., \tilde{w}_N, \tilde{X}_1, ..., \tilde{X}_N, B_1, ..., B_N, \sigma^2, \mu_{\theta, B}, \mu_{\theta, W}, \tau^2, \lambda^2) = \\
\frac{p(\tilde{w}_1, ..., \tilde{w}_N, \tilde{X}_1, ..., \tilde{X}_N | B_1, ..., B_N, \beta_2, \sigma^2, \mu_{\theta, B}, \mu_{\theta, W}, \tau^2, \lambda^2)p(\beta_2 | \xi_2^2)}{\int \int \int p(\tilde{w}_1, ..., \tilde{w}_N, X_1, ..., X_N | B_1, ..., B_N, \beta_2, \sigma^2, \mu_{\theta, B}, \mu_{\theta, W}, \tau^2, \lambda^2)p(\beta_2 | \xi_2^2) d\beta_2}.
\tag{4.20}
\]

The prior on $\beta_2$ is $p(\beta_2) \sim N(\alpha, \xi_2^2)$. In order to determine the posterior of $\beta_2$, I will combine the terms in (4.20) that include $\beta_2$, and complete the square in $\beta_2$,

\[
p(\beta_2 | \tilde{w}_1, ..., \tilde{w}_N, \tilde{X}_1, ..., \tilde{X}_N, B_1, ..., B_N, \sigma^2, \mu_{\theta, B}, \mu_{\theta, W}, \tau^2, \lambda^2) \propto exp \left( \tau^2 \lambda^2 + \sigma^2 \lambda^2 + \sigma^2 \tau^2 + (\xi_2^2 \lambda^2 + \xi_2^2 \tau^2) \sum_j (B_j - \overline{B})^2 \right) \times \\
\left( \beta_2 - \frac{\xi_2^2 \left( \lambda^2 + \tau^2 \right) \sum_j \tilde{w}_j (B_j - \overline{B}) - \tau^2 \sum_j \tilde{X}_j (B_j - \overline{B})}{(\xi_2^2 \lambda^2 + \xi_2^2 \tau^2) \sum_j (B_j - \overline{B})^2 + \tau^2 \lambda^2 + \sigma^2 \lambda^2 + \sigma^2 \tau^2} \right)^2 \\
+ \frac{(\mu_{\theta, W} - \mu_{\theta, B}) \lambda^2 \sum_j B_j (B_j - \overline{B}) - \mu_{\theta, W} \lambda^2 \sum_j (B_j - \overline{B})}{(\xi_2^2 \lambda^2 + \xi_2^2 \tau^2) \sum_j (B_j - \overline{B})^2 + \tau^2 \lambda^2 + \sigma^2 \gamma^2 + \sigma^2 \tau^2} \right)\right)^2.
\]

Thus, $p(\beta_2 | \tilde{w}_1, ..., \tilde{w}_N, \tilde{X}_1, ..., \tilde{X}_N, B_1, ..., B_N, \sigma^2, \mu_{\theta, B}, \mu_{\theta, W}, \tau^2, \lambda^2)$ is normal with mean, $\mu_{\beta_2}$ and variance, $\sigma_{\beta_2}^2$.

\[\Box\]

4.2.1 The Mean and Variance of the Posterior Distribution of $\beta_2$

Because $B_j$ is either 0 or 1, both the mean and variance of the posterior distribution of $\beta_2$ can be further simplified. Starting with the mean, let $\eta$ be the proportion of the sample that is black. Because $B_j$ is either 0 or 1,

\[
\sum_j (B_j - \overline{B}) = 0 \tag{4.21}
\]
\[
\sum_j (B_j - \bar{B}) B_j \over N = \eta(1 - \eta) \quad (4.22)
\]
\[
\sum_j w_j^* (B_j - \bar{B}) \over N = \eta(1 - \eta)(\bar{w}_B - \bar{w}_W) \quad (4.23)
\]
\[
\sum_j X_j^* (B_j - \bar{B}) \over N = \eta(1 - \eta)(\bar{X}_B - \bar{X}_W). \quad (4.24)
\]

Substituting (4.21)–(4.24) into (4.17),
\[
\mu_{\beta_2} = \frac{(\lambda^2 + \tau^2)(\bar{w}_B - \bar{w}_W) - \lambda^2(\mu_{\theta,B} - \mu_{\theta,W}) - \tau^2(\bar{X}_B - \bar{X}_W) + \frac{\alpha(\sigma_2^2 + \sigma_2^2 \lambda^2 + \sigma_2^2 \tau^2)}{\xi^2 N \eta(1 - \eta)}}{\lambda^2 + \tau^2} \quad (4.25)
\]

where \(\bar{w}_B\) is the mean of the dependent variable for blacks and \(\bar{w}_W\) is the mean of the dependent variable for whites. Let \(\bar{X}_B\) and \(\bar{X}_W\) and \(\bar{\theta}_B\) and \(\bar{\theta}_W\) have similar definitions. Note it follows from \(\tilde{X}_1, ..., \tilde{X}_N|\tilde{\theta}_1, ..., \tilde{\theta}_N, \lambda^2 \sim \text{IND} N(\bar{\theta}_j, \lambda^2)\) that \(\bar{X}_B = \bar{\theta}_B\) and \(\bar{X}_W = \bar{\theta}_W\) as \(N \to \infty\). Now when \(N\) is a very large but finite number, (as in a large enough \(N\) for appropriate survey sampling)
\[
\mu_{\beta_2} \approx \frac{(\lambda^2 + \tau^2)(\bar{w}_B - \bar{w}_W) - \lambda^2(\mu_{\theta,B} - \mu_{\theta,W}) - \tau^2(\bar{\theta}_B - \bar{\theta}_W)}{\lambda^2 + \tau^2}. \quad (4.26)
\]

If I am omniscient and know the prior means of the \(\theta\) distribution exactly, then \(\mu_{\theta,B} = \bar{\theta}_B\) and \(\mu_{\theta,W} = \bar{\theta}_W\). Substituting into (4.26), and canceling \(\lambda^2\) and \(\tau^2\) appropriately,
\[
\mu_{\beta_2} \approx (\bar{w}_B - \bar{w}_W) - (\bar{\theta}_B - \bar{\theta}_W), \quad (4.27)
\]

which is the unbiased OLS estimate of \(\beta_2\) when \(\beta_1 = 1\) and \(\theta_j\) is known for all \(j\) and contains no measurement error.

In cases where I am not omniscient, \(\lambda^2\), the measurement error of \(\theta_j\) and \(\tau^2\), the prior variance on \(\theta_j|\text{race}_j\) act as weights. The posterior mean is dependent on the difference of the sample means of the dependent variable and the weighted difference between the prior means of the true ability and the sample means of the observed test score. As in all Bayesian estimates, more data will reduce the possibility of Bayesian shrinkage toward the
4.3. Bias in the Race Coefficient

prior. In the case of the MESE model, more data can come in two forms. First, more data can mean an increase in $N$ which is the number of people who participate in the test. In addition, more data can come in the form of an increase in $M$ or the number of items on the test. As $M$ increases, estimates of $\theta_j$ will be more accurate.

The posterior variance can also be simplified. Substituting (4.21)–(4.24) into (4.18) gives

$$
\sigma_{\beta_2}^2 = \frac{\xi_2^2(\tau^2\lambda^2 + \sigma^2\lambda^2 + \sigma^2r^2)}{(\xi_2^2\lambda^2 + \xi_2^2r^2)\eta(1 - \eta) + \tau^2\lambda^2 + \sigma^2\lambda^2 + \sigma^2r^2},
$$

which goes to 0 as $N \to \infty$ as would be expected.

4.3 Bias in the Race Coefficient in Two Circumstances

In the section above, I derived the mean and variance of the posterior distribution of the $\beta_2$. Using these estimates, I can examine some situations in which bias may result in the posterior estimates. In particular, I look at what happens to the posterior estimate of $\beta_2$, when the prior on $\theta_j$ is not conditioned on race and when the measurement error is high.

4.3.1 The Conditioning Model Revisited

Mislevy (1991) shows that bias will occur in any population statistic if the prior on $\theta_j$ is not conditioned on the correct set of background variables. In the MESE model, bias will occur in the regression coefficients if in the conditioning model (4.1), $\theta_j$ is not conditioned on the covariates in the regression equation. For the case of the black-white wage gaps, this means that the prior on $\theta_j$ must be conditioned on race in order to avoid bias. The following theorem and proof demonstrate the bias that will exist. As noted in Section 4.2, there is no closed form solution to the MESE model, but the MESE simplified model (4.16) gives some intuition as to the bias that will occur.

**Theorem 4.3.1.** In order to avoid bias in the estimate of the coefficients on the covariates, the prior on $\theta_j$ in the MESE model must be conditioned on the covariates included in the regression equation.
Proof. I showed that the unbiased OLS estimate in (4.27) occurs when $N$ is very large but finite and the prior means equal the $\theta$ distribution means exactly. Now I will examine the case when $\mu_{\theta,W} = \mu_{\theta,B}$ or when $\theta_j$ is not conditioned on race. Continue to let $N$ be a finite but very large number. Then, following from (4.26)

$$
\mu_{\beta_2, \text{no cond}} = \frac{(\gamma^2 + \tau^2)(\bar{w}_B - \bar{w}_W) - \tau^2(\bar{X}_B - \bar{X}_W)}{\gamma^2 + \tau^2}.
$$

(4.29)

The unbiased estimate of $\mu_{\beta_2}$ is

$$
(\bar{w}_B - \bar{w}_W) - (\bar{\theta}_B - \bar{\theta}_W).
$$

(4.30)

The estimate of $\mu_{\beta_2, \text{no cond}}$ does not equal the unbiased mean. The bias in the estimate is

$$
bias = \frac{\gamma^2(\bar{X}_B - \bar{X}_W)}{\gamma^2 + \tau^2}.
$$

(4.31)

Because of differential access to quality education, I believe $\bar{X}_B - \bar{X}_W$ will be negative. The bias pushes our estimates of $\beta_2$ in the negative direction insinuating a larger racial wage gap than actually exists.

Thus, $\theta_j$ must be conditioned on race in order to avoid bias in the estimates of the regression coefficients in the MESE model. Take the following example. A labor economist wants to quantify the black-white wage gap and has the following data: the log of wages $w_j$, an indicator variable for race $B_j$ (where $B_j$ is 1 if the person is black and 0 if the person is white) and a measure of human capital that will be an observed test score, $\phi_j$. For simplicity, I consider that the observed test score is from a test that is 1-item long. Thus, the data for the observed test score is really just a scaled version of a right/wrong answer where everyone who answered the item correctly gets the ‘correct’ score and everyone who answered the item incorrectly gets the ‘incorrect’ score.

A standard and reasonable analysis would be to regress $w_j$ on $B_j$ and $\phi_j$. If the estimated coefficient on $B_j$ is negative, a possible conclusion would be that a black-white wage gap existed, even when controlling for cognitive skills. Using simulation, I examine what happens
when an analysis like this is performed.

As in the previous simulations, the truth is

\[
\text{TRUTH} : w_j = 6.0 + 0.2\theta_j - 0.1B_j + \varepsilon_j
\]

\[
\varepsilon_j \sim N(0, \sigma^2).
\]

Let \( N = 280 \) with 140 blacks and 140 whites. Let \( \sigma^2 \) be 0.25 such that \( \varepsilon \) is randomly generated from a \( N(0,0.25) \) distribution. I generate \( \theta_j \) from a \( N(-0.5, 1.25) \) distribution for blacks and a \( N(0.5, 1.25) \) distribution for whites. Each simulation is analyzed 1000 times and the estimates in the table are the means of the 1000 estimates for each simulation.

I generate a test that is 1 item long. The difficulty parameter is set at 0 and the discrimination parameter is set at 1. Note in this simulation, the difficulty of the test item falls directly in the middle of the two means for the two \( \theta_j | \text{race}_j \) distributions. By setting the item parameter in this way, the test item is better at estimating blacks in the upper part of distribution and whites in the lower part of the distribution. Using the 2-PL model, I then randomly generate an item response for each examinee where the item response is drawn from a \( \text{Ber}(P(x_j = 1|\theta_j, a, b)) \) distribution where \( P(x_j = 1|\theta_j, a, b) = \frac{1}{1+\exp(a(\theta_j-b))} \). I scale the test such that the scoring scale and the \( \theta \) scale are equal in order to avoid bias in the estimate of \( \beta_1 \). I determine the true mean of \( \theta \) for everyone who answered the item correctly is 0.85 and the true mean of \( \theta \) for everyone who answered the item incorrectly is \(-0.85\). I set \( \phi_j \) such that a correct answer is equal to 0.85 and a wrong answer is equal to \(-0.85\). I then estimate the regression coefficients when \( w_j \) is regressed on \( B_j \) and \( \phi_j \), my measure of \( \theta_j \). The results from the simulation are shown in Table 4.1 and marked as ELR-No Cond.

The estimated coefficients in the regression equations are decently close to the true coefficients. However, the results are biased too positive for the intercept and biased too negative for the race coefficient. Given that the test score is only 1-item long, I would not expect these results to be exact and might consider these close enough.

Now consider a situation where I use the prior knowledge that on average blacks have lower \( \theta_j \)s than whites. I condition \( \theta_j \) on race. In order to test what happens in this situation,
I repeat the same simulation as in ELR-No Cond: except I condition the observed test score variable on race, resulting in an $\phi_j$ which has four possible scores: black/correct, black/incorrect, white/correct, and white/incorrect instead of just two possible: correct or incorrect. As in ELR-No Cond, the truth is (4.32), $N = 280$ with 140 blacks and 140 whites, $\sigma^2 = 0.25$, $\theta_j$ is generated from a $N(-0.5, 1.25)$ distribution for blacks and a $N(0.5, 1.25)$ distribution for whites and the test is 1 item long with a difficulty parameter of 0 and a discrimination parameter of 1.

Using the 2-PL model, I randomly generate an item response for each examinee where the item response is drawn from a $Ber(P(x_j = 1|\theta_j, a, b))$ distribution where $P(x_j = 1|\theta_j, a, b) = \frac{1}{1+\exp(a(\theta_j-b_j))}$. I scale the test, by determining the true mean $\theta$ for blacks who answered the item correctly is 0.47. I also determine the true mean $\theta$ for whites who answered the item correctly is 1.1. In addition, I determine the true mean $\theta$ for blacks who answered the item incorrectly is $-1.1$ and similarly for whites is $-0.47$. I then set $\phi_j$ such that the score for black/incorrect is $-1.1$ which is less than the score for white/incorrect which is $-0.47$. In addition, the score for black/correct is set to 0.47 and the score for white/correct is set to 1.1. Intuitively, I might guess that results from such an analysis would, in fact, bias the estimates of our regression coefficients. In particular, I might expect that the estimated black-white wage gap would seem to be larger than it was.

In fact, as can be seen in Table 4.1, estimates of the regression coefficients in the simulation where I condition on race called ELR-Cond are much closer to the truth than the estimated coefficients when I do not condition on race. This seem quite shocking. Essentially, the regression coefficient estimates are consistent when using an observed test score that gives a black person a lower score than a white person even when both people answer the test item correctly.

While it seems counterintuitive in the elementary linear regression example, this same concept is behind the idea of conditioning $\theta_j$ on race in the MESE model. In order to further understand what is happening, I look at simulations that compare the estimates of two different, but similar MESE models.

The truth in each simulation is the same as in the ELR simulations, (4.32). Let $N = 280$ with 140 blacks and 140 whites. Let $\sigma^2$ be 0.25 such that $\varepsilon$ is randomly generated from
4.3. BIAS IN THE RACE COEFFICIENT

a $N(0, 0.25)$ distribution. I continue to generate $\theta_j$ from a $N(-0.5, 1.25)$ distribution for blacks and a $N(0.5, 1.25)$ distribution for whites. Each simulation is analyzed 1000 times and the estimates in the table are the means of the 1000 estimates for each simulation.

As in the ELR simulations, I use a 1 item test, where the difficulty parameter is 0 and the discrimination parameter is 1. Using the 2-PL model, I randomly generate an item response for each examinee where the item response is drawn from a $\text{Ber}(P(x_j = 1|\theta_j, a, b))$ distribution where $P(x_j = 1|\theta_j, a, b) = \frac{1}{1 + \exp(a(\theta_j - b))}$.

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). The priors on the regression coefficients are

\begin{align*}
\sigma^2 &\sim Unif(0, 1000) \\
\beta_0 &\sim N(0, 10000) \\
\beta_1 &\sim N(0, 10000) \\
\beta_2 &\sim N(0, 10000).
\end{align*}

In the first simulation, named MESE-No Cond, I set the prior on $\theta_j$ such that it is not conditioned on race: $p(\theta_j) \sim N(0, 1)$. In the second simulation, I condition $\theta_j$ on race and set the priors and hyperpriors as

\begin{align*}
\theta_j|B_j = 1 &\sim N(\mu_{\theta,B}, \frac{1}{\tau_{\theta,B}}) \\
\theta_j|B_j = 0 &\sim N(\mu_{\theta,W}, \frac{1}{\tau_{\theta,W}}) \\
\mu_{\theta,B} &\sim N(0, 1) \\
\mu_{\theta,W} &\sim N(0, 1) \\
\tau_{\theta,B} &\sim \Gamma(1, 1) \\
\tau_{\theta,W} &\sim \Gamma(1, 1).
\end{align*}

The estimates in Table 4.1 demonstrate the need to condition $\theta_j$ on the covariate in the regression. In both the ELR and the MESE models when $\theta_j$ is not conditioned on race,
Table 4.1: Results from Simulations-The Conditioning Model in ELR and MESE models

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\theta$</th>
<th>$B$</th>
<th>$\mu_{\theta,B}$</th>
<th>$\tau_{\theta,B}$</th>
<th>$\mu_{\theta,W}$</th>
<th>$\tau_{\theta,W}$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Truth</strong></td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-0.50</td>
<td>1.25</td>
<td>0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>ELR</strong></td>
<td>Estimate</td>
<td>6.10</td>
<td>0.18</td>
<td>-0.23</td>
<td>-0.20</td>
<td>0.83</td>
<td>0.21</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.02)</td>
<td>(0.07)</td>
<td>(0.02)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.02)</td>
<td>(0.07)</td>
<td>(0.02)</td>
</tr>
<tr>
<td><strong>ELR-Cond</strong></td>
<td>Estimate</td>
<td>5.99</td>
<td>0.21</td>
<td>-0.08</td>
<td>-0.53</td>
<td>0.733</td>
<td>0.53</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.02)</td>
<td>(0.06)</td>
<td>(0.02)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.02)</td>
<td>(0.07)</td>
<td>(0.02)</td>
</tr>
<tr>
<td><strong>MESE</strong></td>
<td>Estimate</td>
<td>6.07</td>
<td>0.31</td>
<td>-0.24</td>
<td>0.01</td>
<td>1.23</td>
<td>0.01</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.04)</td>
<td>(0.15)</td>
<td>(0.04)</td>
<td>(0.14)</td>
<td>(0.87)</td>
<td>(0.14)</td>
<td>(0.87)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.22)</td>
<td>(0.11)</td>
<td>(0.03)</td>
</tr>
<tr>
<td><strong>MESE-Cond</strong></td>
<td>Estimate</td>
<td>6.00</td>
<td>0.27</td>
<td>-0.06</td>
<td>-0.49</td>
<td>1.20</td>
<td>0.47</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.06)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.21)</td>
<td>(0.51)</td>
<td>(0.21)</td>
<td>(0.07)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.17)</td>
<td>(0.15)</td>
<td>(0.16)</td>
<td>(0.04)</td>
</tr>
</tbody>
</table>

Notes: $N = 280$ with 140 whites and 140 blacks. Each simulation run 1000 times. In both ELR and MESE a 1-item test with difficulty parameter equal to 0 and discrimination parameter equal to 1 was used. In ELR-Cond and MESE-Cond, the prior on $\theta_j$ is conditioned on race while in ELR-No Cond and MESE-No Cond, $\theta_j$ is not conditioned on $B_j$. In the ELR specifications, the estimated equation is $w_j = \beta_0 + \beta_1 \phi_j + \beta_2 B_j + \epsilon_j$ where $\phi_j = \theta_j + \nu_j$. $\sigma^2_{\nu} = 0.675$. In the MESE specification, the estimated model is (4.1) - (4.3).

4.3.2 Measurement Error

As noted in Section 4.2.1, Bayesian shrinkage toward the prior will occur when the amount of data is small. It is important to note that the shrinkage will manifest itself in different estimates depending on whether $M$, the number of items or $N$ the number of respondents is small. This section will focus on what happens to the coefficient on race when the number of items is small and the measurement error is high.

In Section 2.5, I demonstrate that the standard errors on $\theta_j$ are dependent on a number of parameters. Standard errors on $\theta_j$ will increase when the number of items is small. In addition, though, when a test has difficulty focused in only one area of the distribution of $\theta_j$, the standard errors for respondents with $\theta_j$ near the difficulty level will be smaller than standard errors on $\theta_j$s outside of the difficulty level of the test. When the data on $\theta_j$
is small (in the sense that the number of items is low or the difficulty level of the test is outside the range of $\theta_j$), the estimates of $\theta_j$ will likely be prone to Bayesian shrinkage to the prior. This will additionally shrink the estimates of the mean and standard deviations of the $\theta$ distributions conditioned on race. This can affect the estimates of the regression coefficients.

Recall that in Section 4.2.1, I showed that the posterior mean of $\beta_2$ (4.26) using the simplified MESE model (4.16) is equal to the unbiased OLS estimator when $N$ is large and $\beta_1 = 1$. Looking at the OLS estimators of the regression coefficients might give some intuition about the direction of the bias when the measurement error is high. Recall that the OLS estimators are:

$$\hat{\beta}_1 = \frac{\hat{\eta}\hat{\sigma}_{\theta w,B} + (1 - \hat{\eta})\hat{\sigma}_{\theta w,W}}{\hat{\eta}\hat{\sigma}_{\theta\theta,B} + (1 - \hat{\eta})\hat{\sigma}_{\theta\theta,W}},$$

and

$$\hat{\beta}_2 = (\bar{w}_B - \bar{w}_W) - \hat{\beta}_1 (\bar{\theta}_B - \bar{\theta}_W).$$

Then, when the measurement error is high, and the mean and standard deviation of $\theta$ are biased due to Bayesian shrinkage, the estimate of $\beta_1$ will be biased. The direction of the bias will depend on the ratio of the bias on the variance and covariances of the two race groups. So, for example, say that $\sigma_{\theta w,B} = \alpha_{\theta w,B} \sigma_{\theta w,B}$, $\sigma_{\theta w,W} = \alpha_{\theta w,W} \sigma_{\theta w,B}$, $\sigma_{\theta\theta,B} = \alpha_{\theta\theta,B} \sigma_{\theta\theta,B}$, and $\sigma_{\theta\theta,W} = \alpha_{\theta\theta,W} \sigma_{\theta\theta,W}$. The estimate of $\beta_1$ is clearly not consistent when the estimates of the variances and covariances that depend on $\theta_j$ are biased. The direction of the bias, though is not as clear. It will be determined by the magnitude of the $\alpha$s.

Additionally, the bias on $\beta_2$ will be determined by the direction of the bias on the $\beta_1$ and the direction of the bias on the difference of the two race groups’ mean $\theta$s. In cases where both $\beta_1$ and the estimated difference $\bar{\theta}_B - \bar{\theta}_W$ are underestimated, $\beta_2$ will be overestimated. Similarly, when $\beta_1$ and the estimated difference $\bar{\theta}_B - \bar{\theta}_W$ are overestimated, $\beta_2$ will be underestimated. When $\beta_1$ is overestimated and the estimated difference $\bar{\theta}_B - \bar{\theta}_W$ is underestimated or vice versa, the direction of the bias on $\beta_2$ will be determined by which is greater.

A simulation study is likely to be helpful in illustrating this point. I compare the results of 2 simulations below. The truth in each simulation is (4.32). Let $N = 280$ with 140
blacks and 140 whites. Let \( \sigma^2 \) be 0.25 such that \( \varepsilon \) is randomly generated from a \( N(0, 0.25) \) distribution. I generate \( \theta_j \) from a \( N(-0.5, 1.25) \) distribution for blacks and a \( N(0.5, 1.25) \) distribution for whites. Each simulation is analyzed 1000 times and the estimates in the table are the means of the 1000 estimates for each simulation.

Then for each simulation, I generate a test. In the first simulation, the test is only 1 item long, so all estimates of \( \theta_j \) will have high measurement error. I randomly generate the difficulty parameter from a \( N(0, 0.1) \) distribution and the discrimination parameter from a \( Unif(0, 2) \) distribution. Note in this simulation, the difficulty of the test item falls directly in the middle of the two means for the two conditional \( \theta \) distributions. In the second simulation the test is 10 items long. I generate the difficulty parameters for the 10 items from a \( N(0, 1) \) distribution and the discrimination parameters from a \( Unif(0, 2) \) distribution. In this simulation, the ten test items are also in the middle of the two distributions, but have a greater variance and so cover a wider range of the distribution of \( \theta \).

Using the 2-PL model, I then randomly generate an item response (or vector of item responses as is the case in the second simulation) for each examinee where the item response is drawn from a \( Ber(P(x_{ij} = 1|\theta_j, a_i, b_i)) \) distribution where 

\[
P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1 + \exp(a_i(\theta_j - b_i))}.
\]

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). The priors on the regression coefficients are stated in (4.33) and the priors on \( \theta_j \) are stated in (4.34). The priors on \( \theta_j \) are conditioned on race in order to avoid the bias demonstrated in Section 4.3.1. In addition, I set the hyperpriors on \( \mu_{\theta,B}, \mu_{\theta,W}, \tau_{\theta,B} \), and \( \tau_{\theta,W} \) such that the prior does not contain any information about whether or not blacks or whites have higher \( \theta_j \). The results from the simulation are shown in Table 4.2. These two simulations show that when the measurement error is high, the resulting estimates are biased, but low measurement error eliminates the bias. In the high measurement error specification, the means of both conditional distributions on \( \theta \) distributions are biased toward zero (as I would expect from a test whose difficulty is lower than the mean for whites and higher than the means for black) and the standard deviations of the distributions are also biased toward zero. This implies that the estimated difference between the mean scores for blacks and whites is also too small. This shrinkage
4.3. BIAS IN THE RACE COEFFICIENT

Table 4.2: Results from Simulations-Increasing Measurement Error

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\theta$</th>
<th>$B$</th>
<th>$\mu_{\theta,B}$</th>
<th>$\tau_{\theta,B}$</th>
<th>$\mu_{\theta,W}$</th>
<th>$\tau_{\theta,W}$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-0.50</td>
<td>1.25</td>
<td>0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td>High Estimate</td>
<td>6.00</td>
<td>0.31</td>
<td>-0.067</td>
<td>-0.38</td>
<td>0.89</td>
<td>0.37</td>
<td>0.90</td>
<td>0.21</td>
</tr>
<tr>
<td>Measurement S. E.</td>
<td>0.07</td>
<td>0.12</td>
<td>0.11</td>
<td>0.22</td>
<td>0.20</td>
<td>0.22</td>
<td>0.20</td>
<td>0.07</td>
</tr>
<tr>
<td>Error</td>
<td>0.066</td>
<td>0.09</td>
<td>0.10</td>
<td>0.20</td>
<td>0.06</td>
<td>0.18</td>
<td>0.08</td>
<td>0.04</td>
</tr>
<tr>
<td>Low Estimate</td>
<td>6.00</td>
<td>0.21</td>
<td>-0.10</td>
<td>-0.49</td>
<td>1.22</td>
<td>0.48</td>
<td>1.22</td>
<td>0.25</td>
</tr>
<tr>
<td>Measurement S. E.</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.11</td>
<td>0.10</td>
<td>0.11</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>Error</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.11</td>
<td>0.10</td>
<td>0.12</td>
<td>0.09</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Notes: $N = 280$ with 140 whites and 140 blacks. Each simulation run 1000 times. In the High Measurement Error specification, the data was generated from a 2PL model with only 1 item with a discrimination parameter generated from a Unif (0, 2) and a difficulty parameter generated from a N(0, 0.1). In the Low Measurement Error specification, the data was generated from a 2PL model with 10 item with discrimination parameters generated from a Unif (0, 2) and difficulty parameters generated from a N(0, 1.0).

in the conditional distributions of $\theta$ results in regression coefficients which are also biased with the coefficient on race biased toward zero and the coefficient on $\theta_j$ biased away from zero. Even though the MESE model accounts for the measurement error, in the case of the high measurement error specification, the test is so uninformative, that the resulting estimates of $\theta_j$ are themselves biased, causing additional bias in the regression coefficients.

In the low measurement error specification, the test’s difficulty is centered at 0 as in the high measurement error specification, but the additional items on the test spread across a larger range of the $\theta$ distribution. Both the increase in the number of items and the wider spread of difficulty parameters produces better estimates of $\theta_j$ for both blacks and whites. The resulting population parameters of $\theta$ are better estimated which results in better regression coefficient estimates. In the low measurement error specification, the estimates of the regression coefficients almost exactly reproduce the truth.

Thus, MESE model produces inconsistent estimates of the regression coefficients when the measurement error is high. When the number of items is small, the estimates of the population parameters of $\theta$ suffer from Bayesian shrinkage. One way to avoid this problem is to ensure that the measurement error is small enough through increasing the number of items, increasing the span of the difficulty parameters among the items, or increasing the discrimination of the items on the test. Recommendations for the length of the test, the variability of the items on the test, and the sample size will follow later in this chapter in
Section 4.5 to help reduce the possibility of this problem.

4.4 Assumptions in the MESE model

In the previous section, I showed two ways in which bias can occur in the coefficient on race: if the measurement error is high, and if the prior on $\theta_j$ is not conditioned on the covariates in the regression equation. In the following section, I discuss two assumptions inherent in the MESE model. First, I examine the assumption about the shape of the prior distribution of $\theta$ and second I examine the assumption about what IRT model to use. I show that assumptions about the shape of the prior distribution on $\theta$ are relatively robust as long as both the true distribution of $\theta$ and the prior distribution on $\theta$ are symmetric. Some bias does occur, though, if the true shape of $\theta$ is skewed but a symmetric prior distribution is assumed. I will also show that misspecification of the IRT model does not greatly affect the estimates of the regression coefficients, though it can affect the estimates of the mean and standard deviation of the $\theta$ distributions.

4.4.1 The Prior Distribution on $\theta$

Throughout the development of the MESE model thus far, I have assumed the prior distribution of $\theta$ is normally distributed. However, Dresher (2006) shows that $\theta$ is not necessarily normally distributed. In order to determine how sensitive the MESE model is to the choice of shape of the prior distribution on $\theta$, I analyze the MESE model in three simulations using different priors and generating distributions for $\theta$.

The truth in each simulation remains (4.32). Let $N = 280$ with 140 blacks and 140 whites. Let $\sigma^2$ be 0.25 such that $\varepsilon$ is randomly generated from a $N(0, 0.25)$ distribution. Each simulation is analyzed 1000 times and the estimates in the table are the means of the 1000 estimates for each simulation.

I generate a ten-item test where the discrimination and difficulty parameters are generated from $Unif(0, 2)$ and $N(0, 0.8)$ distributions respectively. Using the 2-PL model, I randomly generated a vector of item responses for each examinee where an item response is drawn from a $Ber(P(x_{ij} = 1|\theta_j, a_i, b_i))$ distribution where $P(x_{ij} = 1|\theta_j, a_i, b_i) =$
\[ \frac{1}{1+\exp(a_i(\theta_j-b_i))}. \]

In each simulation, \( \theta_j \) is generated from a different distribution which I explain further below.

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). The priors on the regression coefficients are stated in (4.33). The priors on the \( \theta_j \) distribution are different for each simulation and explained below.

In the first simulation, named Normal Data, Uniform Prior, I generate \( \theta_j \) from a \( N(0.5,1.25) \) distribution for whites and a \( N(-0.5,1.25) \) distribution for blacks. The prior on \( \theta_j \) is

\[
\theta_j|B_j = 1 \sim Unif(\theta_{lowB}, \theta_{highB}) \]
\[
\theta_j|B_j = 0 \sim Unif(\theta_{lowW}, \theta_{highW})
\]
\[
\theta_{lowB} \sim Unif(-5,-2)
\]
\[
\theta_{highB} \sim Unif(2,5)
\]
\[
\theta_{lowW} \sim Unif(-5,-2)
\]
\[
\theta_{highW} \sim Unif(2,5).
\]

In the second simulation, named Uniform Data, Normal Prior, I generate \( \theta_j \) from a \( Unif(-1.435,2.435) \) distribution for whites and a \( Unif(-2.435,1.435) \) for blacks. I set the parameters on the uniform distribution such that it has the same mean and standard deviation as the normal distributions used in the Normal Data, Uniform Prior simulation. In this simulation, the priors on \( \theta_j \) are assumed to be normally distribution as stated in (4.34).

In the third simulation, named Skewed Normal Data, Normal Prior I mirrored what Dresher (2006) found. Dresher found a skewness of \(-0.3\) in a sample from the National Assessment of Educational Progress. So, I generate \( \theta_j \) from a skewed normal distribution with a skewness of \(-3\) and mean of \(-0.5\) and a standard deviation of 1.25 for blacks and I generate \( \theta_j \) from a skewed normal distribution with a skewness of \(-3\) and mean of 0.5 and a standard deviation of 1.25 for whites. In this simulation, the priors on \( \theta_j \) are again
## Table 4.3: Results from Simulations-Testing the Shape of the Prior Distribution on $\theta$

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\theta$</th>
<th>$B$</th>
<th>$\mu_{\theta,B}$</th>
<th>$\tau_{\theta,B}$</th>
<th>$\mu_{\theta,W}$</th>
<th>$\tau_{\theta,W}$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Truth</strong></td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-0.50</td>
<td>1.25</td>
<td>0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>Normal Data</strong></td>
<td>Estimate</td>
<td>6.00</td>
<td>0.19</td>
<td>-0.11</td>
<td>-0.31</td>
<td>1.93</td>
<td>0.31</td>
<td>1.93</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.24)</td>
<td>(0.19)</td>
<td>(0.24)</td>
<td>(0.20)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.24)</td>
<td>(0.19)</td>
<td>(0.24)</td>
<td>(0.20)</td>
</tr>
<tr>
<td><strong>Uniform Data</strong></td>
<td>Estimate</td>
<td>6.00</td>
<td>0.19</td>
<td>-0.11</td>
<td>-0.49</td>
<td>1.18</td>
<td>0.49</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.10)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.10)</td>
</tr>
<tr>
<td><strong>Skewed Normal Data</strong></td>
<td>Estimate</td>
<td>6.00</td>
<td>0.23</td>
<td>-0.08</td>
<td>-1.40</td>
<td>0.74</td>
<td>-0.42</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.05)</td>
<td>(0.09)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.07)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.05)</td>
<td>(0.09)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.07)</td>
</tr>
</tbody>
</table>

Notes: N = 280 with 140 whites and 140 blacks. Each simulation run 1000 times. A ten item test with discrimination and difficulty parameters generated from $Unif(0,2)$ and $N(0,0.8)$ distributions respectively was used. In Normal Data, Uniform Prior, $\theta_j$ was generated from a $N(-0.5,1.25)$ distribution for blacks and a $N(0.5,1.25)$ distribution for whites, but the prior on $\theta_j$ was assumed to be uniform. In Uniform Data, Normal Prior, $\theta_j$ was generated from a $Unif(-2.435,1.435)$ distribution for blacks and a $Unif(-1.435,2.435)$ distribution for whites, but the prior on $\theta_j$ was assumed to be normal. In Skewed Normal Data, Normal Prior, $\theta_j$ was generated from a skewed $N(-0.5,1.25)$ distribution with a skewness of $-3$ for blacks and a skewed $N(0.5,1.25)$ distribution with a skewness of $-3$ for whites, but the prior on $\theta_j$ was assumed to be normal.

The results in the first two simulations suggest that there is little bias when the prior and the generating distribution are both symmetric. However, the case in which the data are generated from a non-symmetric distribution, but assume a prior that is symmetric shows more bias. In this latter case, the estimated coefficients on $\theta_j$ and race are 0.23 and $-0.08$ respectively. While these coefficient estimates are not too unreasonable (the estimates when a symmetric prior and data source are used are 0.19 and 0.11), what is more disturbing are the poor estimates of the means and standard deviations for the estimates of $\theta$ for both blacks and whites. Dresher (2006) saw similarly poor estimates of the mean and standard deviation of the distribution of $\theta$ when she assumed a normal prior on $\theta$ whose distribution was actually skewed.
4.4. ASSUMPTIONS IN THE MESE MODEL

When the skewed normal distribution is assumed to be normal, the estimated mean and standard deviation for the distribution of $\theta$ for blacks is $-1.40$ and $0.74$ and the estimated mean and standard deviation for the distribution of $\theta$ for whites is $-0.42$ and $0.78$. However, in the case of the simulation in which the data actually come from a uniform distribution but the prior is assumed to be normal, the estimating mean and standard deviation are quite close to what they should be. Dresher (2006) found that the bias induced by the skewness was less likely to be a problem when a saturated conditioning model was used instead of a small conditioning model. These results suggest that if the estimates of $\theta_j$ are desired, a fuller conditioning model is likely needed. However, if the coefficient estimates are desired, a smaller conditioning model will likely yield acceptable results.

4.4.2 The IRT Model

In many cases, a social scientist will be using a test or assessment that was constructed by others. When this occurs, it is recommended that the researcher use the IRT model employed during the construction, scaling, and scoring of the test. In addition, if IRT parameter estimates are available, those estimates ought to be used, as usually items are used and tested in numerous circumstances and so those parameter estimates will be more accurate. (For example, in the case of the 1992 National Adult Literacy Survey, many of the items on the test were taken from other assessments including the 1985 Young Adult Literacy Assessment and the 1990 Workplace Literacy Survey. Thus, the item parameters were estimated not just with the 24,944 people who participated in the 1992 NALS, but also the people who participated in the 1985 YALA, and the 1990 WLS.)

It is possible, however, that the test will be developed by the social scientist, or that IRT modeling information will not be available. Thus, I examined how misspecification of the IRT model could affect the resulting MESE estimates. I conducted four simulations in which the IRT model was misspecified. In each of the simulations, the generating IRT model is the 2-PL model.

The truth in each simulation is as before (4.32). Let $N = 280$ with 140 blacks and 140 whites. Let $\sigma^2$ be 0.25 such that $\varepsilon$ is randomly generated from a $N(0, 0.25)$ distribution. I generate $\theta_j$ from a $N(-0.5, 1.25)$ distribution for blacks and a $N(0.5, 1.25)$ distribution
for whites. Each simulation is analyzed 1000 times and the estimates in the table are the means of the 1000 estimates for each simulation.

Then for each simulation, I generate a 10 item test, with discrimination parameters generated from a Uniform (0, 2) and difficulty parameters generated from a Normal (0, 1). Using the 2-PL model, I then randomly generate a vector of item responses for each examinee where the item responses are drawn from a Bernoulli distribution with a probability function defined as
\[
P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1+\exp(a_i(\theta_j-b_i))}.
\]

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). The regression coefficient priors are stated in (4.33) and the priors on \(\theta_j\) are stated in (4.34).

In order to test the sensitivity of the MESE model to misspecification of the IRT model, I change the IRT model used in (4.2). Up until now, simulation using the MESE model have always estimated \(\theta_j\) using the generating IRT model with the generating item parameters. However, in the first simulation called Simulation 2PL Data-1PL Sim, I estimate the MESE model using the Rasch model instead of the 2-PL model. I use the true difficulty parameters, but as in the Rasch model, set all discrimination parameters equal to 1. In the second simulation, called 2PL Sim, \(N(0.5, 0.5)\) error on \(a_j\), I model the data with a 2-PL model, but use discrimination parameters, \(a^* = a + t_a\) where \(t_a\) is random error drawn from a \(N(0.5, 0.5)\) distribution. In the third simulation, called 2PL Sim-\(N(0.1, 0.5)\) error on \(b_j\), I use a 2-PL model and the correct discrimination parameters, but use difficulty parameters \(b^* = b + t_b\) where \(t_b\) is random error drawn from a \(N(0.1, 0.5)\) distribution. In the final simulation, called 2PL Sim-\(N(0.1, 0.5)\) error on \(a_j\) and \(b_j\), I use discrimination parameters \(\tilde{a} = a + t_{\tilde{a}}\) and difficulty parameters \(\tilde{b} = b + t_{\tilde{b}}\). The random errors \(t_{\tilde{a}}\) and \(t_{\tilde{b}}\) are both drawn from a \(N(0, 0.1)\) distribution.

The results in Table 4.4 are quite promising. In the specification using the Rasch model instead of the 2-PL model, the estimates are quite close to the truth. The Rasch model overestimates both the means and the standard deviations of the two \(\theta\) distributions, but only slightly. In addition, the Rasch model underestimates the wage gap, but again, very slightly.

The specification where the 2-PL model was used with \(N(0.5, 0.5)\) error on the dis-
4.4. ASSUMPTIONS IN THE MESE MODEL

Table 4.4: Results from Simulations-Testing Misspecification of the IRT Model

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\theta$</th>
<th>$B$</th>
<th>$\mu_{\theta,B}$</th>
<th>$\tau_{\theta,B}$</th>
<th>$\mu_{\theta,W}$</th>
<th>$\tau_{\theta,W}$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-0.50</td>
<td>1.25</td>
<td>0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td>1PL Sim Estimate</td>
<td>6.00 (0.03)</td>
<td>0.20 (0.02)</td>
<td>-0.09</td>
<td>-0.52 (0.13)</td>
<td>1.28 (0.11)</td>
<td>0.52 (0.13)</td>
<td>1.28 (0.11)</td>
<td>0.25 (0.02)</td>
</tr>
<tr>
<td>S. E.</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.19)</td>
<td>(0.21)</td>
<td>(0.21)</td>
<td>(0.21)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>MC Error</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.19)</td>
<td>(0.21)</td>
<td>(0.21)</td>
<td>(0.21)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>2PL Sim Estimate</td>
<td>5.98 (0.03)</td>
<td>0.20 (0.02)</td>
<td>-0.10</td>
<td>-0.42 (0.12)</td>
<td>1.29 (0.10)</td>
<td>0.60 (0.12)</td>
<td>1.28 (0.10)</td>
<td>0.25 (0.01)</td>
</tr>
<tr>
<td>S. E.</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.14)</td>
<td>(0.14)</td>
<td>(0.14)</td>
<td>(0.14)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>MC Error</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.14)</td>
<td>(0.14)</td>
<td>(0.14)</td>
<td>(0.14)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>2PL Sim Estimate</td>
<td>6.00 (0.02)</td>
<td>0.21 (0.02)</td>
<td>-0.10</td>
<td>-0.47 (0.12)</td>
<td>1.22 (0.10)</td>
<td>0.48 (0.12)</td>
<td>1.22 (0.10)</td>
<td>0.25 (0.01)</td>
</tr>
<tr>
<td>S. E.</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.12)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>MC Error</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.12)</td>
<td>(0.01)</td>
</tr>
</tbody>
</table>

Notes: $N = 280$ with 140 whites and 140 blacks. Each simulation run 1000 times. The data was generated from a 2PL model with discrimination parameters generated from a $\text{Unif}(0, 2)$ and difficulty parameters generated from a $\text{N}(0, 1)$. In 1PL sim, $\theta_j$ is estimated using a Rasch model with correct difficulty parameters, but all discrimination parameters assumed to be equal to 1. In 2PL Sim $\text{N}(0, 0.1)$ error on $\alpha_j$ and $\beta_j$, $\text{N}(0, 0.1)$ error was added to both the discrimination and difficulty parameter. In 2PL Sim $\text{N}(0.1, 0.5)$ error on $\beta_j$, $\text{N}(0.1, 0.5)$ error was added to only the difficulty parameter.

The mean of the 10 difficulty parameters is 0, which is below the mean of the distribution of $\theta$s for whites but above the mean of the distribution of $\theta$s for blacks. Thus, more blacks...
will have true $\theta_j$s below the difficulty parameter and so positive error on the discrimination parameter will imply that it was harder for blacks to get the item correct, increasing their estimated $\theta_j$ values. More whites will have true $\theta_j$s above the difficulty parameters, and so positive error on the discrimination parameter will imply that it was easier for whites to get the item correct, resulting in estimated $\theta_j$ values below their true $\theta_j$ values. This bias in the $\theta$ distribution causes bias in the regression coefficient estimates which are slightly biased away from zero.

The specification where the 2-PL model was used with $N(0.1, 0.5)$ error on the difficulty parameter does an excellent job of returning the regression coefficient estimates. It does not estimate the means of the two $\theta$ distributions well. It overestimates the mean for blacks at $-0.42$ instead of $-0.50$ as well as for whites at $0.60$ instead of $0.50$. The boxplots in Figure 4.2 show the differences in the true $\theta_j$ and the estimated $\theta_j$ for each of the four specifications. Note, that the median of the difference in the true score and the estimated $\theta_j$ is 0 for the Rasch model, the model with error on the discrimination parameter, and the model with $N(0,0.1)$ error on both the discrimination and the difficulty parameters. However, the median of the difference when the error on the difficulty parameter is set to $N(0.1,0.5)$ error is $-0.09$. This is to be expected, given that the error added to the
4.5 RECOMMENDED TEST LENGTH AND ITEM VARIABILITY

Figure 4.2: Boxplots of the Theta Residuals for the Four Simulations where IRT Model Misspecification Occurred.

difficulty parameters is positive. Positive error on the difficulty parameters would suggest that the items were more difficult to answer. Answering more difficult items would increase the estimated $\theta_j$'s. Thus, the positive error on the difficulty parameter causes the $\theta_j$'s to be overestimated. However, despite the error in estimating $\theta_j$ in the third specification, the estimated coefficients in the regression equation remain unaffected.

The specification where the 2-PL model was used with $N(0,0.1)$ error on both the discrimination and difficulty parameters also has promising results. Here, the means and standard deviations of the two $\theta$ distributions are underestimated, but again only slightly. The returns to skills are slightly overestimated.

4.5 Recommended Test Length and Item Variability

As noted in Section 4.3.2, high levels of measurement error in the IRT model can result in inconsistent estimates of $\theta_j$. Furthermore, when $\theta_j$ is measured poorly, estimates of the regression coefficients can be biased as well. This subsection is intended to provide some recommendations on the length of the test, and the variability of the items on the test that is needed in order to ensure that the resulting regression coefficients are consistent.
In Chapter 2, I demonstrated that the accuracy of estimates of $\theta_j$ is dependent on three things, the length of the test, the quality of the test items (how discriminating are the test items), and the match between item difficulty and examinee ability. Recall that

$$SE(\hat{\theta}_j) = \frac{1}{\sqrt{I(\theta_j)}},$$

(4.36)

where $I(\theta_j)$ is the information function which is defined as the sum of $I_i(\theta_j)$. For the 2-PL model,

$$I_i(\theta_j) = \frac{a_i^2 \exp(1.7a_i(\theta_j - b_i))(1 - \exp(-1.7a_i(\theta_j - b_i)))^2}{\exp(1.7a_i(\theta_j - b_i))(1 - \exp(-1.7a_i(\theta_j - b_i)))^2}.$$  

(4.37)

In order to better understand how the length of the test and the variability of the item difficulty can affect the standard errors, I estimated the standard errors of $\hat{\theta}_j$ for nine different specifications of the 2-PL model using a 2-factor simulation study where the two factors were test length and difficulty parameter variability.

I set $\theta$ such that it spans from (-4, 4) in intervals of 0.01, giving me 801 $\theta_j$s. I generate a test of test length $tl$ where $tl$ is randomly chosen for each estimation of the model from the sample (1 item, 5 items, or 10 items) with discrimination parameters equal to 1 and difficulty parameters generated from a $N(0, b_{vi})$ where $b_{vi}$ is randomly chosen from the sample (0.1, 0.5, 1.0). I then estimate the standard error for all 801 $\theta_j$s in my sample. I simulate the test 1000 times each time picking from my two factors and calculate the standard errors in each circumstance. This produces a sample of the standard errors for each of the nine specifications. I calculate the median of the sample of standard errors estimated in each specification. Figure 4.3 shows the median standard error for the span of $\theta_j$ for each of the nine specifications. As expected, the greater the variability of the difficulty parameter and the number of items, the lower the standard errors. In addition, all of the tests show the smallest standard errors around 0, since that was the mean of the generated difficulty parameters. What is compelling is how much higher the standard errors are for tests that are only 1 item in length. For all of the tests whose lengths are only 1, the standard errors are above 5 for any $\theta_j$ greater than 2 or less than -2. Interestingly, the test that is 5 items long, but has a variability of 1.0 does almost as well as the test that is 10 items long, but
4.5. **RECOMMENDED TEST LENGTH AND ITEM VARIABILITY**

Figure 4.3: Median Standard Errors for Different Test Lengths and Difficulty Parameter Variability Across Theta.

only has a variability of 0.5. Neither of the tests that is 5 items long with variability less than 1.0 seems to produce reasonable estimates of $\theta_j$ at the extremes. Both have standard errors above 10 for $\theta_j$s greater than 3.5 or less than -3.5.

In addition to estimating the standard errors for the nine specifications, I also analyze the MESE model in each of the specifications. The truth in each case is (4.32). Let $N=280$ with an even number of blacks and whites in each simulation. Let $\sigma^2$ be 0.25 such that $\varepsilon$ is randomly generated from a $N(0, 0.25)$ distribution. I generate $\theta_j$ from a $N(-0.5, 1.25)$ distribution for blacks and a $N(0.5, 1.25)$ distribution for whites. Each simulation is analyzed 1000 times and the estimates in the table are the means of the 1000 estimates for each simulation.

Then for each simulation, I generate a test of length 1, 5 or 10, with discrimination parameters generated from a Unif $(0, 2)$ and difficulty parameters generated from a $N(0, b)$ distribution where $b$ is one of $(0.1, 0.5, 1.0)$. Using the 2-PL model, I then randomly generate a vector of item response (or item responses depending on the length of the test) for each examinee where the item responses are drawn from a $Ber(P(x_{ij} = 1|\theta_j, a_i, b_i))$ distribution where $P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1+exp(a_i(\theta_j-b_i))}$.

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint
Table 4.5: Selected Results from Simulations-Test Length and Variability of Items

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\theta$</th>
<th>$B$</th>
<th>$\mu_{\theta,B}$</th>
<th>$\tau_{\theta,B}$</th>
<th>$\mu_{\theta,W}$</th>
<th>$\tau_{\theta,W}$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Truth</strong></td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-0.50</td>
<td>1.25</td>
<td>0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>1 item</strong></td>
<td>Estimate</td>
<td>6.00</td>
<td>0.30</td>
<td>-0.08</td>
<td>-0.37</td>
<td>0.90</td>
<td>0.35</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.09)</td>
<td>(0.11)</td>
<td>(0.12)</td>
<td>(0.26)</td>
<td>(0.19)</td>
<td>(0.26)</td>
<td>(0.19)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td>(0.11)</td>
<td>(0.36)</td>
<td>(0.09)</td>
<td>(0.37)</td>
<td>(0.08)</td>
</tr>
<tr>
<td><strong>5 item</strong></td>
<td>Estimate</td>
<td>6.00</td>
<td>0.23</td>
<td>-0.08</td>
<td>-0.42</td>
<td>1.16</td>
<td>0.44</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.20)</td>
<td>(0.19)</td>
<td>(0.20)</td>
<td>(0.19)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.21)</td>
<td>(0.15)</td>
<td>(0.21)</td>
<td>(0.16)</td>
</tr>
<tr>
<td><strong>10 item</strong></td>
<td>Estimate</td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-0.48</td>
<td>1.20</td>
<td>0.46</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.12)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.10)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.10)</td>
<td>(0.09)</td>
<td>(0.11)</td>
<td>(0.09)</td>
</tr>
</tbody>
</table>

Notes: $N = 280$ with 140 whites and 140 blacks. Results are selected from a 2-factor simulation experiment. The two factors were variability of the item difficulty parameter (Var) which could be (0.1, 0.5, 1.0), and the number of items (item) which could be (1, 5, 10). The 3-factor simulation experiment was analyzed 2000 times. The discrimination parameter was always drawn from a $Unif(0, 2)$ distribution.

posterior distribution (using WinBUGS software). The regression coefficient priors are stated in (4.33) and the priors on $\theta_j$ are stated in (4.34). I do not show all 9 sets of results but rather show a few key results in Table 4.5.

As expected, the longer test with more variability in the items produces more consistent estimates. Bias exists in the regression coefficients of all three specifications where the test length was only 1; the estimates of the mean and standard deviation of the two conditional distributions of $\theta$ are all biased toward zero. The estimated regression coefficients underestimate the wage gap and overestimate the return to skills. This makes sense as the average difficulty parameter in all of these cases was 0 which is smaller than the average white generated proficiency and larger than the average black generated proficiency. In addition, test lengths of only 1 had very high standard errors for the estimated parameters. An example of results from a 1-item test with item variability of 1.0 are shown in Table 4.5.

In the specification where the test length is 5 and the item variability is 0.5, there is some bias in the estimated regression coefficients, though there is little bias in the estimated regression coefficients when the test length is 5 and the item variability is 1.0. Standard
errors are still high for a test length of only 5 as shown in Table 4.5.

The three specifications where the test length is 10 showed little bias regardless of the variability of the items. Table 4.5 shows an example where the item variability is 0.5.

Based on Figure 4.3 and the results from analyzing the MESE model in all nine specifications, I recommend that the test length and item difficulty variability be chosen so that the standard errors are below 5 for at least 80% of the \( \theta_j \)s in the distribution and below 10 for all of the \( \theta_j \)s in the distribution. Thus, in cases where the \( \theta \) distribution spans from \((-4, 4)\), the test length should be at least 10 with a variability of at least 0.1. If the test length needs to be smaller, it should not go below 5 and then the variability of the items needs to be at least 1.0.

4.6 Comparing the MESE Model to Other Methods

In summary, the MESE model produces estimates of the mean and standard deviation of \( \theta \) as well as estimates of the regression coefficients. I have shown that the estimates of the regression coefficients will be inconsistent when \( \theta_j \) is not conditioned on race and the measurement error is high. In addition, the MESE model is somewhat sensitive to the shape of the prior distribution on \( \theta \). I noted that the MESE model is not very sensitive to misspecification of the IRT model. In addition, I produced recommendations about the variability of the item difficulty parameter and the test length in order to reduce the standard errors and produce consistent regression coefficient estimates.

The MESE model approach addresses a number of the problems noted in Chapter 3. It models the error different for every individual, deals with the identifiability issues in the classic EIV model, and can be used without the need to produce plausible values. This section is intended to compare the models. Simulation is used to demonstrate where the MESE model performs better than the methods introduced in Chapter 3.

4.6.1 MESE Model Compared to Elementary Linear Regression

The MESE approach models the measurement error unlike the typical elementary linear regression model. Simulation results shown in Chapter 3 (named ELR-Corr1) and repeated
Table 4.6: Results from Simulations-ELR versus MESE

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th>( \beta_0 )</th>
<th>( \theta )</th>
<th>( B )</th>
<th>( \mu_{\theta,B} )</th>
<th>( \tau_{\theta,B} )</th>
<th>( \mu_{\theta,W} )</th>
<th>( \tau_{\theta,W} )</th>
<th>( \sigma^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELR-Corr1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td></td>
<td>-0.50</td>
<td>1.25</td>
<td>0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td>S. E.</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC Error</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MESE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td>6.00</td>
<td>0.21</td>
<td>-0.10</td>
<td></td>
<td>-0.48</td>
<td>1.22</td>
<td>0.48</td>
<td>1.21</td>
<td>0.25</td>
</tr>
<tr>
<td>S. E.</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td></td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.09)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>MC Error</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td></td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.01)</td>
</tr>
</tbody>
</table>

Notes: \( N = 280 \) with 140 whites and 140 blacks. Each simulation run 1000 times. \( \theta_j \) was estimated from a randomly generated ten item test with discrimination and difficulty parameters generated from \( \text{Unif}(0,2) \) and \( N(0,0.8) \) distributions respectively. In ELR-Corr1, \( \phi_j = \theta_j + \nu_j \) where \( \nu_j \sim N(0,0.675) \).

In Table 4.6 compare the results of the elementary linear regression model with that of the MESE model. Recall that the truth is

\[
TRUTH : w_j = 6.0 + 0.2\theta_j - 0.1B_j + \varepsilon_j
\]

\[
\varepsilon_j \sim N(0,\sigma^2),
\]

where \( w_j \) is log weekly wages, \( \theta_j \) is the true ability, \( B_j \) is an indicator variable for race that is 1 when the person is black and 0 when the person is white, and \( \varepsilon_j \) is random error. Let \( N = 280 \). I set \( B_j \) such that the first 140 respondents are white and the second 140 respondents are black. Let \( \sigma^2 \) be 0.25 in all the simulations as well, so \( \varepsilon \) is randomly generated from a \( N(0,0.25) \) distribution. I randomly generate \( \theta_j \) from a \( N(-0.5,1.25) \) distribution for blacks and a \( N(0.5,1.25) \) distribution for whites and then calculate \( w \) from the truth above. Each simulation is analyzed 1000 times. The estimates produced in the tables below are the means of the 1000 estimates for each simulation.

In the elementary linear regression model, I also randomly generate an error term for \( \theta_j \) that I call \( \nu_j \) that is generated from a \( N(0,0.675) \). I then calculate \( \phi_j \) by summing \( \theta_j \) and \( \nu_j \). In this simulation, I estimate the regression equation \( w_j = \beta_0 + \beta_1 \phi_j + \beta_2 B_j + \varepsilon_j \).

In the MESE model, I generate a 10 item test where the discrimination parameter is randomly drawn from a \( \text{Unif}(0,2) \) distribution and the difficulty parameter is randomly drawn from a \( N(0,1) \) distribution. Using the 2-PL model, I then randomly generate a vector of item responses for each examinee where the item responses are drawn from a
4.6. COMPARING THE MESE MODEL TO OTHER METHODS

\[ \text{Ber}(P(x_{ij} = 1|\theta_j, a_i, b_i)) \] distribution where \( P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1 + \exp(a_i(\theta_j - b_i))} \). I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). The regression coefficient priors are stated in (4.33) and the priors on \( \theta_j \) are stated in (4.34). I follow Mislevy (1991) and set the priors on \( \theta_j \) such that they are conditioned on race. I estimate the model in (4.1)–(4.3).

Table 4.6 demonstrates that there is bias in both the coefficient on race and the coefficient on \( \theta_j \) when no measurement error is modeled as in ELR-Corr1. However, when an MESE model is used, there is no bias in the estimated coefficients.

4.6.2 MESE Model Compared to Errors In Variables

Unlike the typical errors-in-variables solution where the magnitude of the error is essentially 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. This is good. 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 addition, the MESE model, unlike most classic errors-in-variable solutions, the prior on \( \theta_j \) is motivated by the data-generation process. When a test is scored using an IRT model, but \( \theta_j \) is modeled using classic test theory (as in the EIV model), the data-generation process and the modeling choice are different.

I compared the results from Simulation EIV-prior of Chapter 3 to those when using the MESE model with the same data. The truth is (4.38). Let \( N = 280 \) with 140 respondents who are white and 140 respondents who are black. Let \( \sigma^2 \) be 0.25 in all the simulations as well, meaning that \( \varepsilon \) is randomly generated from a \( N(0, 0.25) \) distribution. I randomly generate \( \theta_j \) from a \( N(-1.0, 1.25) \) distribution for blacks and a \( N(1.0, 1.25) \) distribution for whites and then calculate \( w_j \) from the truth above. Each simulation is analyzed 1000 times. The estimates produced in the tables below are the means of the 1000 estimates for each simulation.
I then generate a 10 item test where the discrimination parameter is randomly drawn from a $Unif(0, 2)$ distribution and the difficulty parameter is randomly drawn from a $N(1, 0.1)$ distribution. The difficulty parameters are set such that in IRT the measurement error on the estimated $\theta_j$s for whites should be smaller than estimated $\theta_j$s for blacks. Using the 2-PL model, I then randomly generate a vector of item responses for each examinee where the item responses are drawn from a $Ber(P(x_{ij} = 1|\theta_j, a_i, b_i))$ distribution where $P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1+\exp(a_i(\theta_j-b_i))}$.

For the errors-in-variables model, I sum the item responses in order to produce a total score for each individual. I set this sum to equal the observed test score, $\phi$. I then estimate the following errors-in-variables regression model,

$$w_j \sim N(\beta_0 + \beta_1 \theta_j + \beta_2 B_j, \sigma_{\epsilon \epsilon}^2)$$
$$\phi_j \sim N(\delta + \Gamma \theta_j, \sigma_{\nu \nu}^2)$$
$$Cor(B, \theta) \neq 0$$
$$E(\epsilon|\phi, B) = 0$$
$$Cov(\nu, \epsilon) = 0.$$ 

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). The regression coefficient priors are stated in (4.33) and the priors on $\theta_j$ are stated in (4.34). The additional priors on $\alpha$, $\sigma_{\nu \nu}$, and $\delta$ are listed below,

$$\delta \sim N(3.27, 0.01)$$
$$\Gamma \sim N(1.45, 0.01)$$
$$\sigma_{\nu \nu}^2 \sim Unif(0, 1000).$$

In the MESE model, I use the 2-PL model with the true item parameters to estimate $\theta_j$. I again use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). The priors are the same as in the EIV model: (4.33) and (4.34). The results are shown in Table 4.7.
### Table 4.7: Results from Simulations-EIV versus MESE

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\theta$</th>
<th>$B$</th>
<th>$\mu_{\theta,B}$</th>
<th>$\tau_{\theta,B}$</th>
<th>$\mu_{\theta,W}$</th>
<th>$\tau_{\theta,W}$</th>
<th>$\sigma_{{\epsilon,\epsilon}}^2$</th>
<th>$\delta$</th>
<th>$\Gamma$</th>
<th>$\sigma_{{\nu,\nu}}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>-1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.25</td>
<td>3.27</td>
<td>1.45</td>
<td>1.4</td>
</tr>
<tr>
<td>EIV</td>
<td>Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. E.</td>
<td>(0.04)</td>
<td>(0.03)</td>
<td>(0.07)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.15)</td>
<td>(0.14)</td>
<td>(0.02)</td>
<td>(0.10)</td>
<td>(0.07)</td>
<td>(0.35)</td>
</tr>
<tr>
<td>MC Error</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.09)</td>
<td>(0.09)</td>
<td>(0.08)</td>
<td>(0.13)</td>
<td>(0.14)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.05)</td>
<td>(0.33)</td>
</tr>
<tr>
<td>MESE</td>
<td>Estimate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. E.</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.05)</td>
<td>(0.13)</td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC Error</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.05)</td>
<td>(0.12)</td>
<td>(0.08)</td>
<td>(0.11)</td>
<td>(0.09)</td>
<td>(0.02)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: N = 280 with 140 whites and 140 blacks. Each simulation run 1000 times. In the MESE model $\theta_j$ was estimated from a 2-PL IRT model. In EIV, $\theta_j$ was estimated from $\phi_j = \delta + \Gamma \theta_j + \nu_j$. The test was a ten item test with discrimination and difficulty parameters generated from $Unif(0,1)$ and $N(1,1)$ distributions respectively, meaning that the test scores of whites should be better estimated than those of blacks.
The MESE model does a better job of estimating $\theta_j$ for both racial groups. The EIV model underestimates the mean and standard deviation for the $\theta$ distribution for blacks, while overestimating the mean and standard deviation for the $\theta$ distribution of whites. The MESE model estimates of the $\theta$ distribution are much closer to the truth, though it does slightly underestimate the mean and standard deviation of $\theta$ for both blacks and whites. Due to this increased bias in the $\theta$ distribution for the EIV model, the regression coefficient on race is overestimated by almost twice what it should be. The MESE model estimate is much closer to the true race coefficient. In both the EIV and the MESE model, the estimates of the coefficient on $\theta_j$ slightly overestimate the return to skills, with the EIV model doing slightly better than the MESE model.

In addition to bias in the estimates of the coefficients and the distribution of $\theta$ for each group, the EIV model has much higher standard errors than the MESE model. This is somewhat expected given the much larger standard errors that exist on estimates of $\theta_j$ for CTT models over IRT models.

### 4.6.3 MESE Model Compared to Marginal Estimation Procedures

The PV Methodology was shown to produce consistent results in Chapter 3. However, 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. PV imputations are not published for many tests (e.g., the AFQT, the GED, and the HSB assessment). 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.

In addition, in theory the MESE model should slightly reduce the standard errors of the estimates. The PV methodology only produces five draws from the posterior distribution and so the Monte Carlo error will be slightly greater in the PV methodology over the MESE model. I compare the results from Simulation PV in Chapter 3 to results using the MESE model with the same data.

As before, the truth is (4.38). Let $N = 280$ with 140 black respondents 140 white respondents. Let $\sigma^2$ be 0.25 in all the simulations as well, meaning that $\varepsilon$ is randomly
4.6. COMPARING THE MESE MODEL TO OTHER METHODS

Table 4.8: Results from Simulations-PV versus MESE

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\theta$</th>
<th>$B$</th>
<th>$\mu_{\theta, B}$</th>
<th>$\tau_{\theta, B}$</th>
<th>$\mu_{\theta, W}$</th>
<th>$\tau_{\theta, W}$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>6.00</td>
<td>0.20</td>
<td>−0.10</td>
<td>−0.50</td>
<td>1.25</td>
<td>0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td>PV</td>
<td>Estimates</td>
<td>6.00</td>
<td>0.21</td>
<td>−0.10</td>
<td>−0.48</td>
<td>1.22</td>
<td>0.48</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.09)</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.09)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>MESE</td>
<td>Estimate</td>
<td>6.00</td>
<td>0.21</td>
<td>−0.10</td>
<td>−0.48</td>
<td>1.22</td>
<td>0.48</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.09)</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.09)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.09)</td>
</tr>
</tbody>
</table>

Notes: N = 280 with 140 whites and 140 blacks. Each simulation run 1000 times. A ten item test with discrimination and difficulty parameters generated from $Unif(0, 2)$ and $N(0, 0.8)$ distributions respectively was used. In PV, the PV Methodology was used with five plausible values drawn from the posterior of $\theta_j$.

generated from a $N(0, 0.25)$ distribution. I randomly generate $\theta_j$ from a $N(-0.5, 1.25)$ distribution for blacks and a $N(0.5, 1.25)$ distribution for whites and then calculate $w$ from the truth above. Each simulation is analyzed 1000 times. The estimates produced in the tables below are the means of the 1000 estimates for each simulation.

I then generate a 10 item test where the discrimination parameter is randomly drawn from a $Unif(0, 2)$ distribution and the difficulty parameter is randomly drawn from a $N(0, 1.0)$ distribution. Using the 2-PL model, I then randomly generate a vector of item responses for each examinee where the item responses are drawn from a $Ber(P(x_{ij} = 1|\theta_j, a_i, b_i))$ distribution where $P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1+\exp(a_i(\theta_j-b_i))}$.

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). The regression coefficient priors are stated in (4.33) and the priors on $\theta_j$ are stated in (4.34).

In the PV simulation, I draw five random imputations from the joint posterior for each individual and then estimate the regression coefficients using the process described in Section 3.2.3. In the MESE model, MCMC is used to sample across the entire $\theta$ distribution in order to estimate the regression coefficients.

The results are shown in Table 4.8. Both the MESE model and the PV model produce estimates of the regression coefficients that are consistent. In addition, both produce estimates of the mean and standard deviations of $\theta$ distributions that are consistent.

Finally, the solution using marginal maximum likelihood procedures currently only mod-
els 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 $\theta_j$ as an independent variable. The MESE model could be varied to use MML estimation instead of MCMC.

Thus, the MESE model performs better than either the typical ELR or EIV models and just as well as the PV model. In addition, the MESE model requires less work of the primary analyst than the PV model.

In order to more fully appreciate the differences in the current methods used and the MESE model, the next chapter will analyze some real world data from the 1992 National Adult Literacy Survey using the various methods available. The dependent variable will be the log of weekly wages and the purpose of the analysis will be to quantify the black-white wage gap in both men and women.
Chapter 5

Using the MESE to Analyze Black-White Wage Gaps

5.1 Current Literature on Black White Wage Gaps

5.1.1 Black-White Wage Gaps

In U.S. labor markets there are large wage disparities between black and white workers. In general, black men earn less than white men, even when comparing groups with similar educational attainment. Altonji and Blank (1999) using data from the March, 1996 Current Population Survey (CPS) show that in 1995 among all workers, white males earned an average hourly wage of $18.96 and annual earnings of $36,169 in comparison to black males, who earned an average hourly wage of $12.41 and annual earnings of $23,645. The basic story is the same when looking at only full-time, full-year workers: white men earn more than black men (Altonji and Blank, 1999). Neal (2005) found similar results: he quotes a black-white ratio of average annual earnings at 0.70 for men who work at least 48 weeks in a year.

Black-white wage gaps in women are significantly smaller than they are for men. And, in some cases, black women have been shown to earn wages similar to or higher than those of white women. In 1995 among all workers, white females earned an average hourly wage of $12.25 and annual earnings of $20,522 in comparison to black females, who earned an
average hourly wage of $10.19 and annual earnings of $17,624 (Altonji and Blank, 1999). The basic story was again the same when looking at only full-time, full-year workers: white women earned more than black women, but the gaps are not as large as they are for men (Altonji and Blank, 1999). Eckstein and Nagypal (2004) suggest that there was wage convergence in black-white wage gaps among women until the 1980s. After the 1980s, this wage gap diverged.

5.1.2 Measuring Human Capital

While labor economists agree that a wage gaps exists between blacks and whites for both men and women, the more pertinent question is why they exist and why are they different by gender. Are the wage gaps a result of discrimination in the work force or are they the result of differing access to pre-market skills?

In order to better understand the underlying causes of the wage gaps, many labor economists use empirical evaluation (i.e., regression based approaches) to understand the gaps. Chapter 3 introduced one such approach: estimate a regression in which the dependent variable is the log wage and independent variables include an indicator variable for race (e.g., 1 if black and 0 if white) along with a series of demographic variables intended to “control for” differing levels of human capital that individuals bring to the labor market. Most commonly these control variables include work experience and education. The central idea is that if, after controlling for observable human capital differences, there remains an estimated race wage gap, one can infer that there is disparate treatment in the labor market. A closely related alternative is to estimate a log wage regression for white workers, again including explanatory variables like experience and education, and then “predict” wages of black workers based on the estimated regression coefficients. These predictions are compared to black workers’ actual wages. If the estimated differences are on average negative, this is taken as evidence of disparate treatment in the labor market. (Altonji and Blank 1999, provide a thorough discussion of these approaches.)

Further introduced in Chapter 3 is the troublesome issue of locating appropriate human capital control variables. In particular, it is common to use years of completed education or a vector of education indicator variables as a means of controlling for human capital. When
using education as the measure of human capital, most studies find that the wage gap drops. (Neal and Johnson, 1996 found unadjusted wage gaps of $-0.244$ and $-0.185$ for men and women respectively. When including a measure of education, the wage gaps dropped to $-0.196$ and $-0.155$, respectively.) Neal (2005) underlines the continued wage gaps that exist even when holding education constant. He additionally notes that black white log wage gaps decline with education level. In particular, Neal suggests that the implied gains from finishing college or high school are “as great or greater for blacks compared to whites in almost every comparison from the 1980-2000 census file” (Neal, 2005, page 23).

But in the U.S. there are massive differences in the quality of education available across communities, and historically blacks and whites have had markedly different access to high-quality education (see, e.g., Card and Krueger, 1992). Lang and Manove (2006) find that overall blacks get approximately 0.75 years less education than do whites. Thus variables that measure only the quantity of education are unlikely to adequately control for racial differences in human capital. Neal and Johnson (1996) suggest that the measurement error induced by blacks getting less education than whites creates a bias toward overstating the black-white wage gap.

In response to this problem, researchers typically follow one of two paths. One path is to collect data with greater detail on education, e.g., course content or specific degrees and college majors (as in Brown and Corcoran, 1997). The second path is to use an independent measure of the cognitive ability as a summary statistic for productive skills one typically develops via schooling and life experience.

Perhaps the most influential example of this latter approach is the work of Neal and Johnson (1996), who analyze the black-white wage gap among young men. In a regression that conditions only on age and ethnicity, the estimated black-white log wage gap for men is $-0.24$. Conditioning on education reduces this gap moderately, to $-0.20$. When instead a measure of cognitive skills (the Armed Forces Qualification Test, AFQT) is included in the regression, the estimated gap declines to $-0.07$. The authors argue that the use of the cognitive test score in their regression allows for more reliable inference about the role of race in labor market outcomes; they conclude that the wage gap primarily reflects a skills gap.
O’Neill (1990) finds similar results to Neal and Johnson (1996). In a regression analysis of male black-white wage gaps that includes both education and AFQT score as independent variables. She shows that while educational attainment has declined over time, educational achievement (as measured by test scores) persists. She concludes that the wage gaps remain because of differing access to the acquisition of human capital due to school quality or family or socioeconomic background.

Lang and Manove (2006) however, disagree. They note that while blacks get less education than whites overall, when conditioning on a measure of cognitive skills (they also use the AFQT) blacks get more education than whites. In an analysis that controls for AFQT score, they show that among men, blacks earn 1.17 more years of education than whites and that among women, blacks earn 1.30 more years than whites. They hypothesize that education is a more valuable signal of productivity for blacks than for whites and so blacks choose to get more education as a way of signaling to potential employers that they are hirable. They further argue that skills and education need to be included as independent variables in a regression analysis that attempts to measure the black-white wage gap. In their analysis of men, when they include only a measure of skills (AFQT score) in the regression equation, they find a black-white wage gap of −0.11. However, when they include both skills and education, they find the black-white wage gap increases to −0.16. They argue this difference in the wage gap occurs as the result of an omitted variables bias, because blacks earn more education than whites conditional on AFQT score.

Lang and Manove (2006) further argue that blacks and whites have differing returns to education. They regress log wage on education, the square of education and interaction terms between race and the education variables. They find that the return to education for blacks and whites are initially lower and then turn more positive as education increases. Their results suggest that the wages of blacks and whites are similar at the tails of the distribution of education, but that blacks earn less than whites in the intermediate levels of education.

Neal and Johnson (1996) and Lang and Manove (2006) are not the only economists using a measure of cognitive skills in their analysis. In fact, there are an increasing number of papers that use a measure of cognitive skill as a variable for helping to understand labor
5.1. CURRENT LITERATURE ON BLACK WHITE WAGE GAPS

market phenomena, like racial wage gaps. Among the tests and assessments used for such purposes are the AFQT (in Neal and Johnson 1996, Lang and Manove, 2006 and other analyses as well), the General Educational Development credential (Tyler, Murnane, and Willett 2000), the National Adult Literacy Survey (Raudenbush and Kasim 1998), and the High School and Beyond assessment (Murnane, Willett, and Levy 1995). Given the growing availability of data on educational achievement, it seems likely that researchers will continue to produce such work.

Chapters 2 and 3 of this thesis however suggest two important lessons for using cognitive test scores as a measure of ability as an explanatory variable in a regression, namely that 1) test scores are measured with error and 2) that the error must be included in the the model or else the resulting estimates of the regression coefficients will be inconsistent. In particular, simulations in Chapter 4 using ELR showed what is commonly found in the EIV literature. There is bias in both the coefficient on race (a bias in the negative direction) and the coefficient on \( \theta_j \) (bias toward zero) when no measurement error was modeled. Since most economists do not deal with the measurement error, this calls into question results from both Neal and Johnson (1996) and Lang and Manove (2006). EIV theory suggests that both Neal and Johnson (1996) and Lang and Manove (2006) may be overstating the wage gap and understating the return to skills.

It is against this backdrop that I conduct my analysis. The rest of this chapter compares several analyses of black-white differences among adults in the U.S. using data from the 1992 National Adult Literacy Survey (NALS). In particular, results from the MESE model described in Chapter 4 are compared with elementary linear regression model results in order to demonstrate the bias that can occur when using an ELR model that does not model the error. In addition, this analysis adds to the growing literature on black-white wage gaps in men and women when an assessment of cognitive skills (functional literacy) is used as a measure of human capital. I discuss how the biased results of the ELR model might provide evidence for policy that eradicates the racial wage gap rather than policy that focuses its attention on eliminating the skills gap.
5.2 The Data: The 1992 National Adult Literacy Survey

The 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. The purpose of the survey is to measure the level of English-language literacy in the United States.

Table 5.1 provides some demographic characteristics of the NALS sample. (Note the tables in this chapter are placed at the end of the chapter for readability because of their length.) 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 both similar educational attainment and similar literacy skills.

The NALS contains 165 items to test the literacy skills of the examinees. While most items require only brief written or oral responses, it was deemed impractical to administer every item to every respondent. Instead, each respondent was randomly administered a booklet designed via the Balanced Incomplete Block (BIB) spiral design which contained a representative sub-sample of approximately one-third of the full set of 165 items.

In a BIB spiral, a test is designed to give each person only a subset of the items. Items are broken into blocks which are then placed in booklets. Each block appears in the same number of booklets and there are enough booklets so that each block appears at least once with every other block. The design is spiraled in that each block appears the same number of times in the front of the booklet, in the middle of the booklet, and in the back of the booklet. It is balanced in that each block appears in the same number of booklets and each booklet is randomly given to approximately the same number of individuals. Finally, it is incomplete in that each individual does not complete all of the items. The table in
5.3. **EMPIRICAL ANALYSIS OF BLACK-WHITE WAGE DIFFERENCES**

The appendix shows the BIB design for the 26 booklets used in the 1992 NALS. While, using a BIB design shortens the test length for each individual, the variability of individual estimates of $\theta_j$ will increase due to the fact that each individual has responded to much fewer items. Thus, when there is a BIB spiral design of the survey, an elementary linear regression model would likely produce even more biased regression coefficients. (For more on the motivation behind the BIB spiraling technique, see Beaton and Zwick, 1992).

Since the primary purpose for conducting the NALS survey was to provide information on the literacy skills of U.S. adults, population estimates were of more interest. Consequently, the NALS data set does not contain individual proficiency estimates, but instead contains five plausible values per content area and individual to aid in calculating population estimates through marginal estimation procedures.

5.3 **Empirical Analysis of Black-White Wage Differences**

Given the data available in the NALS data set, an analysis that estimates log wage regressions is most sensible. In Table 5.2, I present such estimates for two baseline specifications. In both specifications the dependent variable is log weekly wage. 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,” where $potexp = age - yrsschool - 6$. (The number six is included in the potential experience equation because it is 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 I enter potential experience as a quartic. In the second specification, I adopt the standard approach of including also a vector of educational variables.

The study sample in each specification includes only those adults ages 25-55 who work full time. In addition, I follow Neal (2004a and 2004b) by estimating the regressions separately for never married adults (whom I refer to as “single”) and currently married adults. Neal (2004a) argues that marriage market prospects are quite different for black and white men and women and that this generates different patterns of selection into the labor force for women. In particular, Neal notices that married women make up a disproportionate
share of white non-participators, but that among black women, single women are more likely to be out of the labor force. In addition, Neal (2004b) theorizes that married women possess more skills than single women and that among single women, those who are childless possess more skills than those who have children. Neal’s arguments are not without evidence: Zalokar (1990) notes in her report for the United States Commission on Civil Rights, that black husbands earn only two-thirds as much as white husbands earn and that black women contribute one-third of the family income, whereas white women contribute only one-fourth.

I exclude married women entirely from the analysis, because I do not have measures of actual work experience. While “potential experience” is a reasonable approximation of this variable for men and never-married women, it is not a reasonable approximation of work experience for married women. As Neal (2004a) noted, there are large differences in work patterns between married black and white women. So I estimate the regressions in three cases: married men, single men, and single women.

The results in Table 5.2 follow much of what is seen in the literature. 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. As in previous studies, the estimated wage gaps decline modestly when I control for education, $-0.20$ for married men, $-0.08$ for single men, and $-0.05$ for single women, but they remain negative, suggesting a continued wage gap between blacks and whites.

### 5.3.1 Following Neal and Johnson: Using a Measure of Cognitive Skills as the Only Human Capital Measure

Substantial differences are seen when cognitive skills (or functional literacy as measured by the NALS survey) is used as a measure for human capital instead of education. In Table 5.3, I present estimates for four different analyses where in each the wage regression is as follows:

$$
 w_j = \beta_0 + \beta_1 \theta_j + \beta_2 B_j + \beta_3 exp_j^4 + \beta_4 exp_j^3 + \beta_5 exp_j^2 + \beta_6 exp_j + \beta_7 + \epsilon_j
$$

Once again, the dependent variable is log weekly wage, $w_j$. In addition, each specification contains explanatory variables for race, $B_j$—as before an indicator variable equal to 1 if the respondent is black and 0 if the respondent is white—and “potential experience,” $exp_j$ entered as a quartic. The
specifications differ in how $\theta_j$ is entered into the analysis. Recall that the NALS data set does not contain individual proficiency estimates, but instead contains five plausible values per content area and individual to aid in calculating population estimates.

In the first specification, the cognitive test score is the median prose plausible value for each individual. Elementary linear regression is then used to calculate the estimates of the coefficients. Recall from Chapter 3 that this method of using the median plausible value is used by many researchers. In the second specification, the plausible value methodology is applied using the five prose plausible values available in the data set. In the third specification, the MESE model is employed with a prior on $\theta_j$ which is conditioned on race. In the fourth and final specification, the MESE model is again used, however the prior on $\theta_j$ is not conditioned on race. A direct comparison of Neal and Johnson’s (1996) strategy to the MESE is not possible since the NALS does not contain one estimate, however one might argue that the median PV is a Bayes estimate of $\theta_j$. Since I am using the median PV in an elementary linear regression without modeling error as Neal and Johnson do with the AFQT score, this comparison is likely sufficient.

The results in Table 5.3 are not surprising considering the results that were seen in Chapters 2-4. The black-white log wage gap is the most positive when the median PV is used with an elementary linear regression: $-0.05$ for married men, $0.10$ for single men, and $0.13$ for single women. As the simulations in Chapter 3 showed the median PV shrinks the distribution of literacy for each individual. Theory leads us to believe that using the median PV will underestimate the black-white wage gap. The results in Table 5.3 provide further evidence of this phenomenon. Results for the PV methodology and the MESE-Cond analyses are quite similar. The MESE-Cond analysis estimates the wage gap at $-0.08$ for married men, $0.02$ for single men, and $0.01$ for single women. However, the MESE-No Cond analysis overestimates the black-white log wage gaps each subgroup: $-0.12$ for married men, $-0.01$ for single men, and $-0.05$ for single women. Recall the analysis in Chapter 4 that showed the resulting bias that exists when the prior distribution on $\theta_j$ is not conditioned on those variables in the regression analysis.

The MESE results for men are reasonably consistent with Neal and Johnson (1996). 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 0.02 for the single men. For the single male sample I cannot reject the hypothesis that there is no race-based disparate treatment in the labor market. Note additionally, that I hypothesize that Neal and Johnson (1996) would find a smaller black-white wage gap for men if they included the measurement error in their analysis based on the results from the simulation studies performed in Chapter 4.

For the single female sample the estimated coefficient on race (using MESE) is positive, 0.01, and not estimated with enough precision to reject the hypothesis of no disparate treatment in the labor market. 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. (Neal, 2004a and 2004b, provides a discussion of the difficulties in assessing labor market effects of race for women.) Black, Haviland, Sanders, and Taylor (2006), 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.

In the case of comparing the economic return to literacy, comparisons of $\beta_1$ are not ideal, because the distribution of literacy is different for the four specifications. Table 5.3 includes an additional line that compares the increase in observed weekly wages from an increase in one standard deviation of the estimated $\theta$ distribution for that specification. In the case of the estimates using the median plausible value, the resulting increase in the economic return to literacy is only about 9-12%. (Since the dependent variable is log wages, the regression coefficients can be thought of as a change in percent of weekly wages.) Recall that using the median plausible value shrinks the distribution of $\theta$ resulting in a $\theta$ distribution that has a standard deviation which is too small. The PV and MESE models show that the estimated market returns to literacy are substantial. Notice that the returns to literacy are similar for men and women and fall somewhere between 20-27%. As expected, the estimates for the MESE-No Cond model shows economic returns that overestimate those seen in the MESE-Cond model.
5.3. EMPIRICAL ANALYSIS OF BLACK-WHITE WAGE DIFFERENCES

5.3.2 Following Lang and Manove: Including Both Education and a Cognitive Test Score as Measures of Human Capital

Lang and Manove (2006) determined that the racial wage gap increased when both education and a measure of cognitive skills are included in the regression analysis. They argue that this increase in the racial wage gap is an indication of an omitted variable bias. They further argue education should be included in the regression analysis, because years of education is a signal to employers. It is hypothesized by Lang and Manove (2006) that black men get more education conditional on cognitive test score than white men in order to better signal their productivity to potential employers. They conclude that at least some of the black white wage gap among men is a result of differential treatment in the labor market and not entirely a result of differing access to pre-market skills.

However, Chapters 3 and 4 demonstrated that including a variable that is measured with error in an analysis and not accounting for that error can lead to biased results. Recall from Chapter 3, that I demonstrated the well known bias that occurs when there is correlation between the variable measured with error and another covariate in the analysis. In the case of analyses like Lang and Manove, both education and race are highly correlated with cognitive test scores. The lessons from Chapter 3’s simulation results suggest that including education and the cognitive test score in the analysis could bias all of the estimated coefficients.

A simulation analysis shown in Table 5.4 demonstrates this bias. In this simulation, I regress log wages against a test score measured with error, a race identifier, and an education variable which is positively correlated with the test score. The truth is

$$TRUTH : w_j = 6.0 + 0.2\theta_j - 0.1B_j + 0.05Ed_j + \varepsilon_j \quad (5.1)$$

$$\varepsilon_j \sim N(0, \sigma^2),$$

where $w_j$ is log weekly wages, $\theta_j$ is the true ability, $B_j$ is an indicator variable for race that is 1 when the person is black and 0 when the person is white, $Ed_j$ is an education variable that denotes the number of years of schooling for individual $j$, and $\varepsilon_j$ is random error. Let
N = 280 with 140 blacks and 140 whites. Let $\sigma^2$ be 0.25 such that $\varepsilon$ is randomly generated from a $N(0, 0.25)$ distribution. I generate $\theta_j$ from a $N(-0.5, 1.25)$ distribution for blacks and I generate $\theta_j$ from a $N(0.5, 1.25)$ distribution for whites. Each simulation is analyzed 1000 times and the estimates in the table are the means of the 1000 estimates for each simulation.

I generate $Ed_j$ from a Poisson distribution where the mean of the distribution is conditional on race and $\theta_j$. I follow Lang and Manove (2006) and set the education variable so that overall blacks earn about one more year of education than whites conditional on $\theta_j$, but earn about one year less than whites not conditional on $\theta_j$. The regression coefficient estimates for the education variable also came from Lang and Manove (2006). They find an estimated regression coefficient of 0.06 for education and I rounded that down to 0.05 for my truth.

For the specifications using elementary linear regression, random error (generated from a $N(0, 0.675)$) is added to $\theta_j$ to produce an observed test score $\phi_j$ which is used in the analyzed regression equation. In the first specification, ELR-Ed, the estimated equation is $w_j = 6.0 + 0.2\phi_j - 0.1B_j + Ed_j + \varepsilon_j$. In the ELR-No Ed specification, the true wages are still generated from an equation that includes education, but education is not included in the estimated regression equation, $w_j = 6.0 + 0.2\phi_j - 0.1B_j + \varepsilon_j$. This was done to determine how the omitted variable bias interacts with the measurement error bias.

For the MESE specification, I generate a test of ten items. I randomly generate the difficulty parameters from a $N(0, 0.8)$ distribution and the discrimination parameters from a $Unif(0, 2)$ distribution. Note in this simulation, the difficulty of the test item falls directly in the middle of the two means for the two conditional $\theta$ distributions. Using the 2-PL model, I randomly generate an item response vector for each examinee where the item response is drawn from a $Ber(P(x_{ij} = 1|\theta_j, a_i, b_i))$ distribution where $P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1+exp(a_i(\theta_j-b_i))}$.

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). I list the priors for the regression
5.3. EMPIRICAL ANALYSIS OF BLACK-WHITE WAGE DIFFERENCES

coefficients below,

\[ \sigma^2 \sim Unif(0, 1000) \]  \hspace{1cm} (5.2)
\[ \beta_0 \sim N(0, 10000) \]
\[ \beta_1 \sim N(0, 10000) \]
\[ \beta_2 \sim N(0, 10000) \]
\[ \beta_3 \sim N(0, 10000). \]

The priors on \( \theta_j \) are

\[ \theta_j|B_j = 1 \sim N(\mu_{\theta,B}, \frac{1}{\tau_{\theta,B}}) \]  \hspace{1cm} (5.3)
\[ \theta_j|B_j = 0 \sim N(\mu_{\theta,W}, \frac{1}{\tau_{\theta,W}}) \]
\[ \mu_{\theta B} \sim N(0, 1) \]
\[ \mu_{\theta W} \sim N(0, 1) \]
\[ \tau_{\theta B} \sim \Gamma(1, 1) \]
\[ \tau_{\theta W} \sim \Gamma(1, 1). \]

The priors on \( \theta_j \) are conditioned on race following Mislevy (1991). In addition, I set the hyperpriors on \( \mu_{\theta,B}, \mu_{\theta,W}, \tau_{\theta,B}, \) and \( \tau_{\theta,W} \) such that the prior does not contain any information about whether or not blacks or whites have higher \( \theta_j \).

The results in Table 5.4 are really quite surprising. In the ELR-Education specification, as expected, the \( \beta_1 \) coefficient on \( \theta_j \) is biased toward zero. Interestingly the coefficients on education and race are both biased away from zero. The estimate on race is actually biased even further away from zero at \(-0.15\) than the estimate in the simulation found in Chapter 3 when the data was generated from only a test score and a group identifier, \(-0.14\). (See Chapter 3 Table 3.1 for simulation results.) Thus, including an additional variable that is correlated with \( \theta \) (in this case education) actually biases race even further away from zero.

Perhaps even more interesting are the results when education is not included in the ELR model. The results shown in ELR-Omit Education show that (in this case at least)
the measurement error bias and the omitted variable bias seem to almost cancel one another in the estimates on race and test. The only place where bias remains is in the estimate of the intercept.

The MESE-Education specification where error is modeled does a much better job than the ELR-Education. While the coefficient on $\theta_j$ is still slightly biased toward zero and the coefficients on race and education are slightly biased away from zero, none of these biases are large. The lesson from these simulations are that if education is part of the truth, error must be modeled and education must be included in the analysis in order to reduce bias in the regression coefficient estimates.

Since the MESE model does effectively reproduce the true data, I estimated the regression coefficients when education is included in the model using the NALS data. Table 5.5 shows the results in columns (b) for married men, single men, and single women.

The results are quite interesting. The coefficient for education was estimated to be 0.05, 0.03, and 0.05 for married men, single men, and single women respectively. In all three populations, education was statistically significant. My results are similar to those of Lang and Manove (2006) who found an estimate of 0.06 for men (they did not separate married and single men in their analysis.) In addition, $\theta_j$ remained statistically significant, though its coefficient estimate dropped for all three populations. In the case of married men, the coefficient on $\theta_j$ was 0.175 when education was not included and 0.097 when education was included. For single men, the coefficient on $\theta_j$ was 0.197 when education was not included and 0.140 when education was included and for single women, the estimate was 0.199 when education was not included and 0.124 when education was. Lang and Manove found also their estimate on the coefficient in front of $\theta_j$ dropped from 0.26 to 0.15 when they included education in the analysis.

In both the case where education is included in the analysis and when it is not, single people seem to generate the highest returns to literacy skills, followed by married men. Interestingly, single men also have the lowest returns to education with married men and single women have exactly equal returns to education. Single men also have the lowest means for their $\theta$ distributions with single women being the highest. One possible explanation for the higher returns to education for single men is that education is a stronger signal for single
black men than it is for married black men. Since married men, on average, are older than never-married, single men, Lang and Manove’s (2006) signaling theory may be affected by the age of the individual. Future analysts may want to examine Lang and Manove’s signaling theory about education for men and women but separate them by marital status and age groups.

What is perhaps most interesting, though, is what happened to the estimate on the coefficient on race. For married men, like Lang and Manove (2006) our estimate, became more negative (going from \(-0.08\) to \(-0.13\)). In both the case where education was included and when it was not, the coefficient on race is statistically significant suggesting that the racial wage gap among married men cannot be explained entirely by pre-market factors supporting Lang and Manove’s (2006) position. For single men, the coefficient also dropped, going from 0.02 to 0.006, but in both cases I cannot reject the null hypothesis that there is no race-based disparate treatment in the labor market. For single women, however, the coefficient became more positive when I include education in the analysis, going from 0.01 to 0.02. Here again, the null hypothesis cannot be rejected. Thus, in the case of single men and women, there is evidence to support Neal and Johnson’s (1996) position that racial wage gaps are mainly the result of differing access to pre-market skills.

5.3.3 Interaction Effects Between Race and Cognitive Ability

Lang and Manove (2006) further argue that the return to education is different for black and white men. They analyze a regression equation that includes an interaction effect between race and education. In so doing, they find that the return to education is initially lower for blacks and then turns more positive as blacks get more education. They find that the wages for blacks and whites are similar at the tails of the educational distribution, but that blacks earn less than whites in the middle of the distribution. Lang and Manove’s interaction terms are between education and race, but this leads to the question of what happens when an interaction term is included between ability and race. In particular, I am interested in what bias may or may not occur when \(\theta_j\) is not analyzed as having error.

Once again, I start with a simulation experiment. In this simulation, I regress log wages against a test score measured with error, a race identifier, and an interaction term. As in
the case of the simulation involving both education and the test score, I follow Lang and Manove’s lead. I include both the observed test score and the square of the observed test score as well as an interaction between group identifier and the observed test score and an interaction between group identifier and the square of the observed test score. Thus, in this specifications, the simulated data are generated from a model where the truth is

\[ TRUTH : w_j = 6.0 + 0.10\theta_j + 0.5B_j + 0.010\theta_j^2 - 0.15B_j \times \theta_j + 0.50B_j \times \theta_j^2 + \varepsilon_j. \]  

(5.4)

Let \( N = 280 \) with 140 blacks and 140 whites. Let \( \sigma^2 \) be 0.25 such that \( \varepsilon \) is randomly generated from a \( N(0, 0.25) \) distribution. I generate \( \theta_j \) from a \( N(-0.5, 1.25) \) distribution for blacks and a \( N(0.5, 1.25) \) distribution for whites. Each simulation is analyzed 1000 times and the estimates in the table are the means of the 1000 estimates for each simulation.

For the specification using elementary linear regression, random error (generated from a \( N(0, 0.675) \)) is added to \( \theta_j \) to produce an observed test score \( \phi_j \). The estimated equation in the simulation is: \( w_j = 6.0 + 0.1\phi_j + 0.5B_j + 0.010\phi_j^2 - 0.15B_j \times \phi_j + 0.50B_j \times \phi_j^2 + \varepsilon_j \). For the MESE specification, I generate a test of ten items. I randomly generate the difficulty parameters from a \( N(0, 0.8) \) distribution and the discrimination parameters from a \( Unif(0, 2) \) distribution. Note in this simulation, the difficulty of the test item falls directly in the middle of the two means for the two conditional \( \theta \) distributions. Using the 2-PL model, I randomly generate an item response vector for each examinee where the item response is drawn from a \( Ber(P(x_{ij} = 1|\theta_j, a_i, b_i)) \) distribution where

\[ P(x_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1+exp(a_i(\theta_j-b_i))}. \]

I use Markov Chain Monte Carlo (MCMC) machinery to numerically calculate the joint posterior distribution (using WinBUGS software). I use the priors in (5.2) and (5.3). The priors on \( \theta_j \) are conditioned on race. In addition, I set the hyperpriors on \( \mu_{\theta,B}, \mu_{\theta,W}, \tau_{\theta,B}, \) and \( \tau_{\theta,W} \) such that the prior does not contain any information about whether or not blacks or whites have higher \( \theta_j \).

The results in Table 5.6 are not too surprising. As expected, in the elementary linear regression specification where no error is modeled, the estimated coefficients on \( \theta_j \) and \( \theta_j^2 \) are underestimated at 0.08 and 0.006 instead of 0.10 and 0.010 respectively. The coefficient
on group identifier (which in this case was positive) is estimated away from zero. In addition, the estimates on the two interaction terms are also biased, with the interaction term between group identifier and \( \theta_j \) being biased in the negative direction away from zero and the interaction term between group identifier and \( \theta_j^2 \) being biased in the negative direction toward zero.

The results from the MESE model, however, are much closer to what they should be. Here, the interaction terms are slightly biased away from zero. Thus, once again modeling the error is necessary in order to produce consistent regression coefficient estimates.

In order to better understand Lang and Manove’s hypothesis about interaction terms, I used the NALS data to test the hypothesis that the return to literacy is different between blacks and whites for married men, single men and single women. Returning to Table 5.5, the results in column (c) suggest that the return to skills for blacks and whites is likely not very different. In all three cases, (married men, single men, and single women) there was no statistical significance for any of the interaction terms. This is similar to what Neal and Johnson (1996) found where there was no statistical significance in interaction terms among AFQT score and race. In addition, Neal and Johnson (1996) examined the marginal effect of AFQT on log wages for each racial group separately and found no statistically significant differences.

When examining the black-white log wage gap, results when including the interaction term are similar to those when no interaction term is included, estimating the wage gap at \(-0.11\) for married men, \(-0.05\) for single men, and 0.09 for single women. As was the case when no interaction term was used, I cannot reject the null hypothesis of no racial wage gap for single men and women.

In comparing the economic returns to literacy, none of the squared terms of the test score are statistically significant either which is also in-line with what Neal and Johnson (1996) found. This suggests that the returns to skills for blacks and whites and men and women are likely not quadratic.
5.3.4 Residual Analysis

In order to examine the fit of the MESE model, I examine the residuals \((w_{\text{obs}} - \hat{w})\) where \(w_{\text{obs}}\) are the observed wages and \(\hat{w}\) are the predicted wages. Since \(\theta_j\) is a latent variable that is unobserved, calculating \(\hat{w}\) is not obvious. In order to determine a residual of the model, I estimated a posterior distribution of \(\text{resid}_j\) for each \(j\) where \(\text{resid}_j = w_j - (\beta_0 + \beta_1\theta_j + \beta_2B_j + \beta_3\exp_j + \beta_4\exp_j^2 + \beta_5\exp_j^3 + \beta_6\exp_j^4)\). I took the mean of this posterior distribution as the estimate of the unstandardized residuals, \(\text{resid}_j\). I then calculated the studentized residual by dividing \(\text{resid}_j\) by the standard deviation of all of the posterior means of \(\text{resid}\).

I then drew four diagnostic plots for each of the MESE model analyses fit in this chapter: a plot of the studentized residuals, a QQ plot that compares the distribution of the studentized residuals with the standard normal distribution, a plot of the studentized residuals by fitted value, and a boxplot of the studentized residuals by race. Each of these plots is located in the appendix for each of the nine analyses whose results are in Table 5.5.

Overall, the diagnostic plots do not vary much by specification (no education or interaction term, an education term, or an interaction term) for each population. In all nine cases, the plot of the studentized residuals is well scattered around 0. Since there is no pattern to the residuals, nonconstant variance is not indicated in the plot. Note that in each of the nine specifications, \(\sigma^2\) is estimated between 0.53 and 0.57. In the case of married men and single women, there is one studentized residual that is a great outlier that is equal to \(-10\). Further investigation of this data point shows that in both cases, this person’s wages were reported as 0, despite being in labeled as in the labor force. This person is probably mislabeled in the data. Rerunning the analysis without this data point did not change anything significantly.

The diagnostic plots also do not indicate any nonlinearity. In both the scatter plot of the residuals and the scatter plot of the residuals against the fitted values, there is no curved trend. This suggests that the log transformation of the wage is likely appropriate.

All of the QQ-plots show some deviation from the standard normal at the tails with married men having the least deviation. This is not unexpected given that \(N\) in the married men is three times the \(N\) in the case of both single men and single women. Finally, the
5.4 Discussion

Over the past decades there has been an explosion in the availability of cognitive test score data, in nationally representative samples drawn for social science research, in data collected for the purpose of measuring students’ progression in school, and data used to examine adult ability at the national and international level. More and more social scientists are turning to this kind of data for conducting further analysis.

This chapter provides a relevant cautionary tale. I show that one of the commonly employed practices for using a cognitive skill measure as an explanatory variable—simply placing the test score in an elementary linear regression—can lead to serious misunderstanding. I demonstrate the use of the MESE model in practice and show results from the model that are similar to results found using the PV methodology. This comparison of results is possible, because the data I use is from the 1992 National Adult Literacy Survey which provides five plausible values in its data set. However, in the case of many other similar data sets (i.e., the AFQT) plausible values are not available. Thus, my MESE model would be an ideal model to be used in such a case rather than the oft-used method of linear regression.

Accounting for measurement error in the measure of cognitive skills using the MESE model instead of an ELR model is not the only difference between work like that of Neal and Johnson (1996), and Lang and Manove (2006) and my work. Because I use a data set different from that of Neal and Johnson and Lang and Manove, interpretation of the results will be slightly different in two ways.

First, the 1992 NALS is a test intended to measure functional English-language literacy. It includes such items as determining the amount of money to be left for a tip and completing order forms to buy something through a mail catalogue (Kirsch, et al., 2000). The AFQT is an achievement test of verbal and mathematical skill. While I can find no papers that directly compare the two tests, the AFQT is likely a harder test than the NALS. The AFQT is used by the military to determine to test individuals for fitness for military service. One
might argue that the AFQT is a better measure of the skills needed in the workplace. However, the basic skills that are tested on the NALS are likely useful in any job setting, whereas highly specific knowledge and skills are less versatile and do not travel well to other jobs. In addition, no one test can measure the skills needed in every workplace. Using various tests and assessments adds to the growing body of research around the role of skills on earnings. Finally, the results I obtain in my analysis are quite similar to those of Neal and Johnson (1996) and Lang and Manove (2006) suggesting that while the two tests may measure different skill sets, skills in general play a large role in reducing the wage gaps.

My results must also be interpreted differently from Neal and Johnson (1996) and Lang and Manove (2006) due to the timing of the collection of the measure of skills. In the NLSY data, skills are measured when the individuals are 16-18 years of age and the wage data is determined approximately 10 years later. Neal and Johnson (1996) argue that their measure of skills is pre-market which is advantageous for identifying causation. Suppose that white workers are more likely to obtain high-level jobs because of discrimination, and that these high-level jobs utilize and reinforce skills to a greater extent than low-level jobs. If this were the case, an analysis that uses contemporaneous data on skills and wages would show that controlling for skills greatly reduces the black-white earnings gap despite the role of discrimination in determining who was hired for the high-level jobs.

However, skills change over time and not all changes in skills are the result of racial discrimination. Estimates using data like that of the NLSY can only be interpreted as the portion of the overall racial wage gap attributable to human capital formation before the age of 16. Data that has contemporaneous measures of skills and wages, though looks at the role of current human capital that includes both pre-market and in-market skills. It is interesting to see what portion of the current racial wage gap is attributable to current human capital. As noted above, the results of Neal and Johnson (1996) and Lang and Manove (2006) do not differ substantial from my results suggesting that skill disparities regardless of when they are measured play an important role in the the racial wage gap.

This chapter then adds to a growing body of evidence that most of the black-white disparity in wages in U.S. labor markets is the consequence of racial disparity in skills. In particular, in regressions that adjust only for potential experience, I estimate black-
white log wage gaps of $-0.34$ for married men, $-0.18$ for single men, and $-0.20$ for single women. My estimates from the MESE model suggest that after accounting for measured differences in literacy the gaps are $-0.08$, $0.02$, and $0.01$ respectively. Even when accounting for differences in literacy and education in order to account for different possible education signals to potential employers, the gaps are $-0.13$, $0.01$, and $0.02$. respectively.

It is interesting to note that the wage gap remains for married men even when conditioning on skills, but that no wage gap remains for either single women or single men. There is a large literature in labor economics that studies the “marriage pay premium.” In these studies, labor economists have long noted that married men earn substantially more than men who are not currently married. These cross-sectional wage differentials persist when controls are introduced for education, race, region, age, or work experience, and even occupation and industry. Korenman and Neumark(1991) report differentials in the 10-40 percent range. Nakosteen and Zimmer (1987) suggest that one possible explanation for the marriage pay premium is that the skills that are valued in the marriage market are also valued in the labor market. Thus one explanation for the difference in the racial wage differentials among single and married men in the NALS has to do with the marriage pay premium. Being single may affect wages even more than being black. Further research that controls for age and looks at an interaction affect between marital status and race could explore this hypothesis.

Regardless of whether or not single men endure discrimination, it is reasonable to infer that most, though quite likely not all, of the race wage gap stems from the skills gap. My results therefore underscore the importance of policies that effectively address this skills gap prior to entering the labor market.
### Table 5.1: Sample Characteristics of the NALS

<table>
<thead>
<tr>
<th></th>
<th>Black Men</th>
<th>Black Women</th>
<th>White Men</th>
<th>White Women</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>1665</td>
<td>2807</td>
<td>7449</td>
<td>9404</td>
</tr>
<tr>
<td><strong>Avg. Age</strong></td>
<td>39.4</td>
<td>39.4</td>
<td>40.5</td>
<td>42.4</td>
</tr>
<tr>
<td><strong>Marital Status</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion never married</td>
<td>0.39</td>
<td>0.38</td>
<td>0.28</td>
<td>0.19</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion Still in HS</td>
<td>0.06</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Proportion &lt; HS</td>
<td>0.29</td>
<td>0.29</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>Proportion HS</td>
<td>0.29</td>
<td>0.29</td>
<td>0.27</td>
<td>0.30</td>
</tr>
<tr>
<td>Proportion &lt; College</td>
<td>0.25</td>
<td>0.29</td>
<td>0.30</td>
<td>0.33</td>
</tr>
<tr>
<td>Proportion College +</td>
<td>0.11</td>
<td>0.09</td>
<td>0.26</td>
<td>0.21</td>
</tr>
<tr>
<td><strong>Literacy Skills</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median Prose PV</td>
<td>−0.55</td>
<td>−0.39</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>Earnings—Full-Time Workers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. Weekly Wage</td>
<td>452.3</td>
<td>397.5</td>
<td>674.6</td>
<td>440.9</td>
</tr>
</tbody>
</table>

Notes: Author’s calculations, National Adult Literacy Survey.
### Table 5.2: Baseline Regressions Using Elementary Linear Regression

<table>
<thead>
<tr>
<th></th>
<th>Married Men</th>
<th>Single Men</th>
<th>Single Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a)</td>
<td>(b)</td>
<td>(a)</td>
</tr>
<tr>
<td>Race</td>
<td>-0.341</td>
<td>-0.204</td>
<td>-0.184</td>
</tr>
<tr>
<td></td>
<td>(0.047)</td>
<td>(0.042)</td>
<td>(0.065)</td>
</tr>
<tr>
<td>Education:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; HS</td>
<td>-0.935</td>
<td>-1.288</td>
<td>-2.132</td>
</tr>
<tr>
<td></td>
<td>(0.087)</td>
<td>(0.189)</td>
<td>(0.226)</td>
</tr>
<tr>
<td>Some HS</td>
<td>-0.735</td>
<td>-0.690</td>
<td>-0.629</td>
</tr>
<tr>
<td></td>
<td>(0.049)</td>
<td>(0.114)</td>
<td>(0.131)</td>
</tr>
<tr>
<td>HS</td>
<td>-0.469</td>
<td>-0.454</td>
<td>-0.501</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(0.068)</td>
<td>(0.071)</td>
</tr>
<tr>
<td>GED</td>
<td>-0.660</td>
<td>-0.646</td>
<td>-0.381</td>
</tr>
<tr>
<td></td>
<td>(0.062)</td>
<td>(0.170)</td>
<td>(0.144)</td>
</tr>
<tr>
<td>&lt;2 yrs Col.</td>
<td>-0.329</td>
<td>-0.179</td>
<td>-0.239</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.077)</td>
<td>(0.073)</td>
</tr>
<tr>
<td>&gt;2 yrs Col.</td>
<td>-0.214</td>
<td>-0.119</td>
<td>-0.165</td>
</tr>
<tr>
<td></td>
<td>(0.047)</td>
<td>(0.087)</td>
<td>(0.098)</td>
</tr>
<tr>
<td>&gt; College</td>
<td>0.270</td>
<td>0.191</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.077)</td>
<td>(0.069)</td>
</tr>
</tbody>
</table>

| adj. $R^2$     | 0.03        | 0.27       | 0.04         | 0.26         |
| N              | 2557        | 710        | 640          |

Notes: All regressions also control for potential experience entered as a quartic. The omitted variable in the education category is College degree. Included in the sample for the regressions are individuals aged 25-55, who work full-time and reported wages and who answered at least one literacy item.
Table 5.3: Regressions Comparing Med PV, PV, and 2 Specifications of the MESE Model

<table>
<thead>
<tr>
<th></th>
<th>Married Men</th>
<th>Single Men</th>
<th>Single Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ELR PV</td>
<td>MESE Cond</td>
<td>MESE No Cond</td>
</tr>
<tr>
<td></td>
<td>Median PV</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Race</td>
<td>-0.054</td>
<td>(0.046)</td>
<td>-0.088</td>
</tr>
<tr>
<td>Lit. Skills:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median PV</td>
<td>0.267</td>
<td>(0.013)</td>
<td>0.274</td>
</tr>
<tr>
<td>NCES Jackknife</td>
<td>0.235</td>
<td>(0.015)</td>
<td>0.243</td>
</tr>
<tr>
<td>MESE-Cond</td>
<td>0.175</td>
<td>(0.011)</td>
<td>0.197</td>
</tr>
<tr>
<td>MESE-No Cond</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increase for 1 SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increase of $\theta$:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.088</td>
<td>0.245</td>
<td>0.212</td>
</tr>
<tr>
<td>White</td>
<td>0.096</td>
<td>0.214</td>
<td>0.205</td>
</tr>
<tr>
<td>Both races</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_{\theta, Black}$</td>
<td>-0.241</td>
<td>-0.237</td>
<td>0.125</td>
</tr>
<tr>
<td>$\mu_{\theta, White}$</td>
<td>0.847</td>
<td>0.849</td>
<td>1.59</td>
</tr>
<tr>
<td>$\tau_{\theta, Black}$</td>
<td>0.330</td>
<td>1.042</td>
<td>1.162</td>
</tr>
<tr>
<td>$\tau_{\theta, White}$</td>
<td>0.360</td>
<td>0.999</td>
<td>1.116</td>
</tr>
<tr>
<td>$\mu_{\theta}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau_{\theta}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>5093</td>
<td>5142</td>
<td>78080</td>
</tr>
<tr>
<td>N</td>
<td>2557</td>
<td>2557</td>
<td>710</td>
</tr>
</tbody>
</table>

Notes: Included in the sample are individuals aged 25-55, who work full-time, reported wages and who answered at least one literacy item.
Table 5.4: Results from Simulations-ELR and MESE Models with Education and $\theta$

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\theta$</th>
<th>$B$</th>
<th>Ed</th>
<th>$\mu_{\theta,B}$</th>
<th>$\tau_{\theta,B}$</th>
<th>$\mu_{\theta,W}$</th>
<th>$\tau_{\theta,W}$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>6.00</td>
<td>0.20</td>
<td>-0.10</td>
<td>0.05</td>
<td>0.50</td>
<td>1.25</td>
<td>-0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td>ELR</td>
<td>Estimate</td>
<td>5.94</td>
<td>0.15</td>
<td>-0.15</td>
<td>0.06</td>
<td>-0.49</td>
<td>1.4</td>
<td>0.50</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.06)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.005)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.12)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td>MC Err.</td>
<td>(0.06)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.12)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>ELR</td>
<td>Estimate</td>
<td>6.55</td>
<td>0.21</td>
<td>-0.11</td>
<td>-0.50</td>
<td>1.4</td>
<td>0.49</td>
<td>1.4</td>
<td>0.28</td>
</tr>
<tr>
<td>Omit Ed</td>
<td>S. E.</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td>MC Err.</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.09)</td>
<td>(0.11)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>MESE</td>
<td>Estimate</td>
<td>5.92</td>
<td>0.19</td>
<td>-0.11</td>
<td>0.06</td>
<td>-0.48</td>
<td>1.22</td>
<td>0.48</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.04)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.004)</td>
<td>(0.12)</td>
<td>(0.12)</td>
<td>(0.12)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td>MC Err.</td>
<td>(0.06)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.004)</td>
<td>(0.11)</td>
<td>(0.09)</td>
<td>(0.11)</td>
<td>(0.09)</td>
</tr>
</tbody>
</table>

Notes: $N = 280$ with 140 whites and 140 blacks. Each simulation run 1000 times. Ten item test.

In the MESE specification, $\theta_j$ was estimated from a 2-PL model. In each of the ELR specifications, $\phi_j = \theta_j + \nu_j$ where $\nu_j \sim N(0,0.675)$. Education was generated to be positively correlated with $\theta$ such that blacks earned 1 year more education when conditioned on $\theta_j$ and 1 year less education when there was no conditioning. In ELR-Omit Ed, the true data was generated with education, but the estimated regression equation left education out.
Table 5.5: Regressions Comparing 3 Specifications of the MESE Model–Including just $\theta$, including $\theta$ and Ed and Including and Interaction between $\theta$ and Race

<table>
<thead>
<tr>
<th></th>
<th>Married Men</th>
<th>Single Men</th>
<th>Single Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta$</td>
<td>$\theta$ &amp; Ed</td>
<td>Interaction</td>
</tr>
<tr>
<td>Race</td>
<td>-0.080 (0.040)</td>
<td>-0.125 (0.038)</td>
<td>-0.105 (0.049)</td>
</tr>
<tr>
<td>Lit. Skills</td>
<td>0.175 (0.011)</td>
<td>0.097 (0.012)</td>
<td>0.178 (0.026)</td>
</tr>
<tr>
<td>Lit. Skills$^2$</td>
<td>0.00036 (0.0088)</td>
<td>-0.0024 (0.017)</td>
<td>-0.027 (0.019)</td>
</tr>
<tr>
<td>Race-Lit.</td>
<td>-0.0280 (0.041)</td>
<td>0.041 (0.068)</td>
<td>0.122 (0.076)</td>
</tr>
<tr>
<td>Race-Lit.$^2$</td>
<td>0.0249 (0.025)</td>
<td>0.0432 (0.031)</td>
<td>0.0121 (0.035)</td>
</tr>
<tr>
<td>Years Educ</td>
<td>0.045 (0.0032)</td>
<td>0.033 (0.0072)</td>
<td>0.045 (0.0079)</td>
</tr>
<tr>
<td>$\mu_{\theta,Black}$</td>
<td>0.125 (0.077)</td>
<td>0.125 (0.074)</td>
<td>0.124 (0.079)</td>
</tr>
<tr>
<td>$\mu_{\theta,White}$</td>
<td>1.59 (0.027)</td>
<td>1.59 (0.025)</td>
<td>1.59 (0.026)</td>
</tr>
<tr>
<td>$\tau_{\theta,Black}$</td>
<td>1.16 (0.065)</td>
<td>1.16 (0.064)</td>
<td>1.16 (0.063)</td>
</tr>
<tr>
<td>$\tau_{\theta,White}$</td>
<td>1.12 (0.021)</td>
<td>1.12 (0.021)</td>
<td>1.11 (0.021)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.545 (0.0078)</td>
<td>0.530 (0.0077)</td>
<td>0.544 (0.0081)</td>
</tr>
<tr>
<td>AIC</td>
<td>78080</td>
<td>77870</td>
<td>78020</td>
</tr>
<tr>
<td>N</td>
<td>2557</td>
<td>710</td>
<td>640</td>
</tr>
</tbody>
</table>

Notes: Included in the sample are individuals aged 25-55, who work full-time, reported wages and who answered at least one literacy item.
Table 5.6: Results from Simulations-ELR and MESE Models when There is an Interaction Term Between Test Score and Race

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Test</th>
<th>Race</th>
<th>Test²</th>
<th>Test×Race</th>
<th>Test²×Race</th>
<th>μθ,B</th>
<th>τθ,B</th>
<th>μθ,W</th>
<th>τθ,W</th>
<th>σ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>6.00</td>
<td>0.10</td>
<td>0.50</td>
<td>0.010</td>
<td>−0.15</td>
<td>0.50</td>
<td>−0.50</td>
<td>1.25</td>
<td>0.50</td>
<td>1.25</td>
<td>0.25</td>
</tr>
<tr>
<td>ELR</td>
<td>Estimate</td>
<td>6.02</td>
<td>0.08</td>
<td>0.68</td>
<td>0.006</td>
<td>−0.21</td>
<td>0.30</td>
<td>−0.50</td>
<td>1.4</td>
<td>0.50</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.06)</td>
<td>(0.04)</td>
<td>(0.08)</td>
<td>(0.02)</td>
<td>(0.05)</td>
<td>(0.02)</td>
<td>(0.04)</td>
<td>(0.09)</td>
<td>(0.09)</td>
<td>(0.07)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.007)</td>
<td>(0.004)</td>
<td>(0.08)</td>
<td>(0.002)</td>
<td>(0.07)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.03)</td>
<td>(0.09)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>MESE</td>
<td>Estimate</td>
<td>6.00</td>
<td>0.10</td>
<td>0.51</td>
<td>0.01</td>
<td>−0.17</td>
<td>0.51</td>
<td>−0.49</td>
<td>1.26</td>
<td>0.48</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>S. E.</td>
<td>(0.005)</td>
<td>(0.004)</td>
<td>(0.009)</td>
<td>(0.002)</td>
<td>(0.06)</td>
<td>(0.03)</td>
<td>(0.09)</td>
<td>(0.03)</td>
<td>(0.09)</td>
<td>(0.002)</td>
</tr>
<tr>
<td></td>
<td>MC Error</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.01)</td>
<td>(0.003)</td>
<td>(0.09)</td>
<td>(0.13)</td>
<td>(0.13)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.003)</td>
</tr>
</tbody>
</table>

Notes: N = 280 with 140 whites and 140 blacks. Each simulation run 1000 times. The truth was copied from Lang and Manove (2006). θj was estimated from a randomly generated ten item test with discrimination and difficulty parameters generated from Unif(0, 2) and N(0, 1) distributions respectively.
### 5.5 Appendix

Table 5.7: BIB Design of the 26 Booklets Used in the 1992 NALS

<table>
<thead>
<tr>
<th>Booklet Number</th>
<th>Block Numbers Contained in Booklet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Core 1 2 13</td>
</tr>
<tr>
<td>2</td>
<td>Core 2 3 9</td>
</tr>
<tr>
<td>3</td>
<td>Core 3 4 7</td>
</tr>
<tr>
<td>4</td>
<td>Core 4 13 8</td>
</tr>
<tr>
<td>5</td>
<td>Core 13 9 6</td>
</tr>
<tr>
<td>6</td>
<td>Core 9 7 10</td>
</tr>
<tr>
<td>7</td>
<td>Core 7 8 11</td>
</tr>
<tr>
<td>8</td>
<td>Core 8 6 12</td>
</tr>
<tr>
<td>9</td>
<td>Core 6 10 5</td>
</tr>
<tr>
<td>10</td>
<td>Core 10 11 1</td>
</tr>
<tr>
<td>11</td>
<td>Core 11 12 2</td>
</tr>
<tr>
<td>12</td>
<td>Core 12 5 3</td>
</tr>
<tr>
<td>13</td>
<td>Core 5 1 4</td>
</tr>
<tr>
<td>14</td>
<td>Core 1 3 8</td>
</tr>
<tr>
<td>15</td>
<td>Core 2 4 6</td>
</tr>
<tr>
<td>16</td>
<td>Core 3 13 10</td>
</tr>
<tr>
<td>17</td>
<td>Core 4 9 11</td>
</tr>
<tr>
<td>18</td>
<td>Core 13 7 12</td>
</tr>
<tr>
<td>19</td>
<td>Core 9 8 5</td>
</tr>
<tr>
<td>20</td>
<td>Core 7 6 1</td>
</tr>
<tr>
<td>21</td>
<td>Core 8 10 2</td>
</tr>
<tr>
<td>22</td>
<td>Core 6 11 3</td>
</tr>
<tr>
<td>23</td>
<td>Core 10 12 4</td>
</tr>
<tr>
<td>24</td>
<td>Core 11 5 13</td>
</tr>
<tr>
<td>25</td>
<td>Core 12 1 9</td>
</tr>
<tr>
<td>26</td>
<td>Core 5 2 7</td>
</tr>
</tbody>
</table>

Notes: BIB design from the National Adult Literacy Survey.
5.5. APPENDIX

Diagnostic Plots

Figure 5.1: Diagnostic Plots of the Studentized Residuals for the MESE model for Married Men with no Education or Interaction Term. Plot (a) is a plot of the studentized residuals. Plot (b) is a QQ plot comparing the distribution of the Studentized Residuals with the Standard Normal. Plot (c) is a plot of the Studentized Residuals by Fitted Values. Plot (d) is a boxplot of the Studentized Residuals by Race.
Figure 5.2: Diagnostic Plots of the Studentized Residuals for the MESE model for Married Men with an Education Variable. Plot (a) is a plot of the studentized residuals. Plot (b) is a QQ plot comparing the distribution of the Studentized Residuals with the Standard Normal. Plot (c) is a plot of the Studentized Residuals by Fitted Values. Plot (d) is a boxplot of the Studentized Residuals by Race.
Figure 5.3: Diagnostic Plots of the Studentized Residuals for the MESE model for Married Men with an Interaction Term. Plot (a) is a plot of the studentized residuals. Plot (b) is a QQ plot comparing the distribution of the Studentized Residuals with the Standard Normal. Plot (c) is a plot of the Studentized Residuals by Fitted Values. Plot (d) is a boxplot of the Studentized Residuals by Race.
CHAPTER 5. USING THE MESE TO ANALYZE BLACK-WHITE WAGE GAPS

Figure 5.4: Diagnostic Plots of the Studentized Residuals for the MESE model for Single Men with no Education or Interaction Term. Plot (a) is a plot of the studentized residuals. Plot (b) is a QQ plot comparing the distribution of the Studentized Residuals with the Standard Normal. Plot (c) is a plot of the Studentized Residuals by Fitted Values. Plot (d) is a boxplot of the Studentized Residuals by Race.
Figure 5.5: Diagnostic Plots of the Studentized Residuals for the MESE model for Single Men with an Education Variable. Plot (a) is a plot of the studentized residuals. Plot (b) is a QQ plot comparing the distribution of the Studentized Residuals with the Standard Normal. Plot (c) is a plot of the Studentized Residuals by Fitted Values. Plot (d) is a boxplot of the Studentized Residuals by Race.
Figure 5.6: Diagnostic Plots of the Studentized Residuals for the MESE model for Single Men with an Interaction Term. Plot (a) is a plot of the studentized residuals. Plot (b) is a QQ plot comparing the distribution of the Studentized Residuals with the Standard Normal. Plot (c) is a plot of the Studentized Residuals by Fitted Values. Plot (d) is a boxplot of the Studentized Residuals by Race.
Figure 5.7: Diagnostic Plots of the Studentized Residuals for the MESE model for Single Women with no Education or Interaction Term. Plot (a) is a plot of the studentized residuals. Plot (b) is a QQ plot comparing the distribution of the Studentized Residuals with the Standard Normal. Plot (c) is a plot of the Studentized Residuals by Fitted Values. Plot (d) is a boxplot of the Studentized Residuals by Race.
Figure 5.8: Diagnostic Plots of the Studentized Residuals for the MESE model for Single Women with an Education Variable. Plot (a) is a plot of the studentized residuals. Plot (b) is a QQ plot comparing the distribution of the Studentized Residuals with the Standard Normal. Plot (c) is a plot of the Studentized Residuals by Fitted Values. Plot (d) is a boxplot of the Studentized Residuals by Race.
Figure 5.9: Diagnostic Plots of the Studentized Residuals for the MESE model for Single Women with an Interaction Term. Plot (a) is a plot of the studentized residuals. Plot (b) is a QQ plot comparing the distribution of the Studentized Residuals with the Standard Normal. Plot (c) is a plot of the Studentized Residuals by Fitted Values. Plot (d) is a boxplot of the Studentized Residuals by Race.
Chapter 6

Discussion and Future Work

6.1 Summary and Contributions

Throughout this dissertation, I examine the issue of measurement error associated with using cognitive test scores as independent variables in regression analyses. In Chapter 2, I introduce two bodies of theory in psychometrics: classic test theory and item response theory and note their advantages and limitations. In Chapter 3, I show that the current methods of modeling the error all have limitations. I show that not modeling the error at all can lead to biased estimates of the regression coefficients. Errors-in-variables analysis cannot model heteroskedastic measurement error and has the typical problem of being unidentifiable. Plausible value methodology can be problematic when used with longitudinal datasets or split sample design data and requires a primary analyst to produce plausible values before a secondary analyst can use the data.

I focus my attention on the social science issue of black-white wage gaps in married men, never-married men, and never-married women. However, as I note in Chapter 1, many areas of social science use cognitive test scores as a measure for ability. It is quite likely that not modeling the error in the test score could be problematic in these analyses as well. For example, I note that educators control for ability when trying to determine the role of parental resources on rates of college participation. A possible simplified model that
would examine this would be,

$$CP_j = \beta_0 + \beta_1 \theta_j + \beta_2 PR_j + \varepsilon_j$$

$$\varepsilon_j \sim N(0, \sigma^2),$$

where $CP_j$ is the log of the observed participation rate for individual $j$, $\theta_j$ is the unobserved latent ability trait, and $PR_j$ is an indicator variable that is 1 when a parent’s resources are low and 0 when a parent’s resources are high. This model looks almost exactly like model (3.1) in Chapter 3. I noted that when $\phi_j$ is substituted for $\theta_j$ in such a model that bias will occur in the regression coefficients. Since I expect that $\theta$ and parental resources are negatively correlated, I would expect that using an elementary linear regression that does not model the error in $\theta_j$ would produce estimate of $\beta_1$ that are biased toward 0 and estimates of $\beta_2$ that are biased away from 0. These results might suggest that the gap in college participation rates among the children of those with high parental resources and those with low parental resources is due less to a skills gap and more to a resource gap, when in fact, the skills gap plays a larger role. There are numerous other examples in the social science literature where not modeling the measurement error can lead to biased results which might lead to incorrect inferences.

A model that can appropriately account for the measurement error in the test scores is desired. In Chapter 4, I take up the matter of finding a model that can solve the measurement error problem using Item Response Theory to model the error. The MESE model described in Chapter 4 does just that. The MESE model has greater flexibility in modeling ability with heteroscedastic measurement error, handles the identifiability issues of the EIV model, can be used with a wider variety of study designs, and does not require any work by a primary analyst.

The MESE model has its own set of limitations as well. I note that the model does not produce consistent results when the prior on $\theta_j$ does not include the other regression covariates. I also note that high amounts of measurement error also produce biased results. I note that the MESE model is somewhat sensitive to the choice of the prior on $\theta_j$, especially when the true distribution of $\theta$ is skewed but the prior assumes it to be symmetric.
6.1. SUMMARY AND CONTRIBUTIONS

One of the strengths of the MESE model is that it is not particularly sensitive to misspecification of the IRT model. In addition, when measurement error is low, the MESE model produces consistent results.

This dissertation provides a relevant cautionary tale. I provide theoretical and empirical arguments that raise concerns about common practices in which researchers use available cognitive skill measures as explanatory variables in empirical work. In addition, I provide a viable alternative approach to using cognitive skills measures in empirical work, the MESE model.

In Chapter 5, I turn my attention to the evaluation of black-white differentials in labor markets. I compare analysis results from an elementary linear regression model that uses the median PV score as the test score to results using the MESE model. I find that the results using the median PV lead to serious misunderstandings. When I accurately account for the measurement error in the test score by using the MESE model, I find results that add to a growing body of evidence. Namely, that most of the black-white disparity in wages in U.S. labor markets is the consequence of racial disparity in skills. For both single men and single women, there is no remaining wage gap, once I control for both experience and cognitive skills. For married men, the wage gap reduces significantly from -0.34 when no controls are in place to -0.08 when I control for experience and skills.

Additionally, I examine the hypotheses of Lang and Manove (2006) who cast doubt on the consensus that the wage gaps are primarily caused by pre-market factors and not labor market factors. They postulate that blacks earn more education conditional on test score, because it is a more valuable signal to employers than it is for whites. Lang and Manove (2006) believe that education and test scores must be included as covariates in the regression analysis, because without education, an omitted variable bias occurs in the coefficient on race. In an analysis of the wage gaps that includes both education and race, I continue to find no remaining wage gap for single men and single women. The wage gap for married men, however, does increase from -0.08 to -0.12. I also find that the return to education is statistically significant (though small). My results support the hypothesis that education is a signal to employers, though I suggest that the signal may be stronger by age or marital status. And except for married men, my results continue to support the consensus that the
wage gaps are primarily caused by a skills gap.

I also examine the possibility of differing returns to skills for blacks and whites. My results indicate that the return to skills for married men, single men, and single women is approximately 20-27% for both blacks and whites. I find no evidence that the return to skills is dissimilar by race.

While I cannot make any definite conclusions based on my results about the cause of the continued wage gaps among blacks and whites in the labor force, my results do suggest the importance of policies that effectively address the black-white skills gap.

6.2 Future Work

The MESE model described in Chapter 4 is as a step in the right direction to better modeling of the measurement error of the test score in regression analyses. Further work in better understanding the model’s behavior in various situations is needed. In particular, there are three areas in which future work could focus methodologically: the identifiability of the model, the expansion of the MESE model with other types of dependent variables and the usability of the model for other researchers.

Identifiability of the Model

While the issue of the identifiability of the model was touched on, further work needs to be done in this area. In Chapter 4, I note that the MESE model handles the identifiability issues of the EIV model, but introduces identifiability issues of its own, through the IRT model embedded in the MESE model. I note that the identifiability issues of the IRT model can be solved by setting the scale of $\theta$. In addition, by analyzing the MESE model in a Bayesian framework, I can set the priors of the MESE model such that the unidentifiability of the IRT model is solved.

Further work in this area includes a more formal proof of the way in which the MESE model solves the identifiability issues of the EIV model. In addition, a formal proof showing that setting the scale of $\theta$ solves the unidentifiability of the MESE model is needed.
6.2. FUTURE WORK

Expansion of the Model

The substantive work in this dissertation focused on the issue of black-white wage gaps where linear regression is one of the more common methods of analysis. However, there are other sets of response variables and other types of latent constructs that the MESE model could be expanded to analyze.

First, the MESE model could be expanded such that it could analyze non-normal response variables. For example, in the labor economic arena, the issues of non-participation and unemployment in the workforce is one of great interest (particularly for women). An expansion of the MESE model to allow for other generalized linear models to be used (e.g., logistic regression) would be useful. This would mainly involve a change in the final level of the hierarchical model.

An additional response variable that would involve an expansion of the model would be another variable measured with error. For example, in the education arena, there are studies which attempt to predict the scores on one test using the scores of another. The MESE model provides an excellent framework from which to expand to handle the non-random error in the response variable. Fox and Glas (2001) and Patz and Junker (1999) might provide some insight into how to expand the MESE model in such a way.

Second, the MESE model could be expanded to analyze other latent constructs. For example, Heckman, Stixrud, and Urzua (2006) examine the effects of both cognitive and non-cognitive abilities (like motivation, self-esteem, and control) on labor market outcomes. In many cases, these non-cognitive abilities are measured using tests. An IRT model is often used to “score” the non-cognitive ability. The MESE model could be easily expanded so that non-cognitive traits or any latent construct measured through the use of IRT can be placed as independent variables in a regression equation.

Usability of the Model

In addition, I argue that the PV methodology is difficult for the average social scientist to use, because the average social scientist does not apply the methodology correctly. In its current form, the MESE model requires quite a bit of statistical know-how from the social
scientist. Most social scientists are not trained in Markov Chain Monte Carlo methods. In order to be more useful to the typical social scientist, future work on the MESE model could include writing the model into a software program that would require the researcher to input the data, the model specifications, and choices on the prior, and then produce appropriate output for analysis and consideration. Once the MESE model has been tested in more places, this write-up could be useful to many other researchers.

In addition to future work in the methodological area, there is continued work to be done in the economic arena. There are three substantive areas that could be explored more fully: 1) how non-participants can be included in the model, 2) the effect of child-rearing choices on wage gaps in women, and 3) wage gaps among other minority groups and using other data sets.

Non-participants and Unemployment

Butler and Heckman (1977) (and Brown, 1984, and Chandra, 2003) note that black men are overrepresented in the group of people opting out of the labor force. In addition, black men are overrepresented in the group of people who are unemployed. A common theory in the literature states that among men, the non-participants and the unemployed have lower skill, and/or motivation, ambition, and effort. This further suggests that the black-white wage gap in men when non-participants and the unemployed are included in the analysis should be considerably larger than analyses that only include full-time workers.

Unlike men who have different labor force participation rates by race, black and white women have similar labor force participation and employment rates. Further research however (Neal 2004a), shows that among lower-skilled women, blacks are relatively more likely to be out of the labor force, while among relatively higher-skilled women, the reverse is true, suggesting that not including non-participant women in an analysis can also underestimate the wage gap. Thus, further work should explore a way in which to include non-participants and the unemployed. Chandra (2003) provides a nice example of how to include non-participants non-parametrically. Further work should look to expand the MESE model such that a non-parametric approach may be possible.
6.2. FUTURE WORK

Child-Rearing Choices

Neal (2004a) provides some compelling initial arguments about the way in which family structure can affect labor force choices, particularly for women. I followed Neal’s arguments and separated my analyses by both gender and marital status. In addition, I did not include married women in this analysis, because my potential experience variable is not a good estimate of actual experience for married women. But if marital status can have such an effect on wage gaps, it is possible that child-rearing choices will also have an effect on wage gaps.

In particular, child-rearing choices may have a large effect on the choices of whether or not women are in the labor force. Black women are more likely to be never married and raising children than white women. Further exploration into the way in which this affects the choices of women to be in the labor force by race and thus affects the measured black-white wage gap should be completed. The NALS data set does provide information about the number of people in a household and how many of those individuals who are under 18, so the analysis is possible.

Other Groups and Other Data Sets

Finally, the results of this dissertation provide only one snapshot of the black-white wage gap for men and women in 1992. Further exploration of the wage gap should include the use of other data sets. In particular, it would be interesting to apply the MESE model to the NLSY dataset that both Neal and Johnson (1996) and Lang and Manove (2006) use in order to determine just how much the measurement error affect their results. (I am currently still in talks with NORC and the DOD who are in charge of the NLSY and ASVAB data to try to get the item responses for the AFQT.)

In addition, further exploration of wage gaps should include other minority groups (e.g. Hispanics) to understand the wage gaps that exists outside of blacks and whites.

Finally, Blau and Kahn (2000) use the International Adult Literacy Survey (IALS) to examine the differences in labor market inequalities internationally. The IALS dataset is quite similar to the NALS dataset. Using the MESE model to examine such issues would
be an easy expansion of the model with the IALS data.
References


Group.

Testing Service.

ernment Printing Office.

Zimowski, M., Muraki, E., Mislevy, R. J. and Bock, R. D. (2006), \textit{BILOG-MG3 for Win-
dows}, Mooresville, IN: Scientific Software.