FALSE POSITIVES IN MULTIPLE
REGRESSION: HIGHLIGHTING THE CONSEQUENCES OF
MEASUREMENT ERROR IN THE INDEPENDENT VARIABLES

by

BENJAMIN ROGERS SHEAR

A.B., Dartmouth College, 2006

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate Studies

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2012

© Benjamin Rogers Shear, 2012
Abstract

Type I error rates in multiple regression, and hence the chance for false positive research findings in the literature, can be drastically inflated when the analyses include independent variables measured with error. Although the bias caused by random measurement error in multiple regression is widely recognized, there has been little discussion of the impact on hypothesis tests outside of the statistical literature. The primary purpose of this thesis is to raise awareness of the problem among methodologists and researchers by demonstrating, in a non-technical manner, the nature and extent of the inflation in Type I error rates for educational and psychological research contexts. This thesis uses computer simulations to demonstrate that, for commonly encountered scenarios, the Type I error rate in a multiple regression model where the independent variables are correlated and measured with random error can approach 1.0, even if the nominal Type I error rate is 0.05. Because nearly all quantitative data in educational and psychological research contain some level of random measurement error, and because multiple regression is one of the most widely used data analytic techniques, this problem should be a serious concern for methodologists and applied researchers. The most important factors causing the problem are summarized, and the implications for research and pedagogy are discussed.
Table of Contents

Abstract ......................................................................................................................... ii
Table of Contents ........................................................................................................... iii
List of Tables ................................................................................................................... v
List of Figures ................................................................................................................ vi
List of Symbols .............................................................................................................. vii
Acknowledgements ...................................................................................................... viii

1 Introduction .............................................................................................................. 1
  1.1 Measurement Error and Hypothesis Testing in Regression ................................. 1
  1.2 Motivating Examples ............................................................................................. 1
  1.3 Rationale ................................................................................................................ 4
  1.4 Study Purpose ......................................................................................................... 5
  1.5 Outline of the Thesis .............................................................................................. 5

2 Background and Literature Review .......................................................................... 6
  2.1 Chapter Summary .................................................................................................... 6
  2.2 Measurement Error and Reliability ....................................................................... 6
  2.3 Motivating Example, Notation and Definitions ..................................................... 9
  2.4 Review of Psychometric and Statistical Literature ............................................. 14
      2.4.1 Attenuation of Correlations ............................................................................ 14
      2.4.2 Bias and Type I Error Rates in Simple Regression ....................................... 15
      2.4.3 Bias in Multiple Regression ......................................................................... 17
      2.4.4 Type I Error Rates in Multiple Regression .................................................. 21
  2.5 An Unrecognized Problem ..................................................................................... 27
  2.6 Purpose of the Study .............................................................................................. 28

3 Study 1: Reviewing and Demonstrating the Impact of Measurement Error in Simple Regression ............................................................................................................. 30
  3.1 Introduction .......................................................................................................... 30
  3.2 Methods ................................................................................................................ 30
      3.2.1 Study Design .................................................................................................. 30
      3.2.2 Data Generation ............................................................................................. 31
      3.2.3 Outcome Variables ....................................................................................... 32
      3.2.4 Data Analysis ............................................................................................... 32
  3.3 Results ................................................................................................................... 33

4 Study 2: False Positives in Multiple Regression Studies when Independent Variables are Measured with Error .......................................................................................... 39
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>39</td>
</tr>
<tr>
<td>4.2</td>
<td>Methods</td>
<td>40</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Study Design</td>
<td>40</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Data Generation</td>
<td>44</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Outcome Variables</td>
<td>45</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Data Analysis</td>
<td>46</td>
</tr>
<tr>
<td>4.3</td>
<td>Results</td>
<td>49</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Part I Results</td>
<td>49</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Part II Results</td>
<td>55</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Part III Results</td>
<td>62</td>
</tr>
<tr>
<td>5</td>
<td>Conclusion</td>
<td>67</td>
</tr>
<tr>
<td>5.1</td>
<td>Study 1 Synopsis</td>
<td>67</td>
</tr>
<tr>
<td>5.2</td>
<td>Study 2 Synopsis</td>
<td>68</td>
</tr>
<tr>
<td>5.3</td>
<td>Implications for Researchers</td>
<td>71</td>
</tr>
<tr>
<td>5.4</td>
<td>Potential Solutions</td>
<td>76</td>
</tr>
<tr>
<td>5.5</td>
<td>Implications for Teaching</td>
<td>77</td>
</tr>
<tr>
<td>5.6</td>
<td>Novel Contributions</td>
<td>82</td>
</tr>
<tr>
<td>5.7</td>
<td>Limitations and Future Directions</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td><strong>Bibliography</strong></td>
<td>86</td>
</tr>
</tbody>
</table>
List of Tables

Table 1. Type I Error Rates in Simple Regression. .................................................. 34
Table 2. Observed Sample Estimates in Simple Regression with Measurement Error. .. 34
Table 3. Summary of Study 2 Design Conditions for Parts I-III. ................................. 43
Table 4. Summary Statistics for 5,000 Experimental Conditions in Part I .................... 49
Table 5. Variable Ordering Based on Linear Model in Part I ...................................... 51
Table 6. Variable Ordering Based on Logistic Model in Part I ..................................... 53
Table 7. Summary Statistics for 2,000 Experimental Conditions in Part II .................... 56
Table 8. Variable Ordering Based on Linear Model in Part II ..................................... 58
Table 9. Variable Ordering Based on Logistic Model in Part II ................................... 59
Table 10. Observed Type I Error Rates in Part III ...................................................... 63
Table 11. Mean Estimated Standardized Coefficient in Part III ................................... 65
Table 12. Mean Estimated Change in R-square in Part III ........................................ 66
List of Figures

Figure 1. Graphical Depiction of the Relationship Among True Scores. ...................... 11
Figure 2. Estimated Standard Errors in Simple Regression with Measurement Error. .... 37
Figure 3. Variability of Simple Regression Slopes with Measurement Error. ............. 38
Figure 4. Histogram of Observed Type I Error Rates in Part I................................. 50
Figure 5. Simple Slopes for Linear Model in Part I............................................. 54
Figure 6. Simple Slopes for Logistic Model in Part I........................................... 55
Figure 7. Histogram of Observed Type I Error Rates in Part II.............................. 56
Figure 8. Observed Type I Error Rates by Reliability of \( u \) and \( R \)-square in Part II..... 61
Figure 9. Schematic of Inferences in Regression..................................................... 81
List of Symbols

Cor(x, y): correlation between two random variables, x and y.

Var(x): the variance of a random variable x.

rel(x): reliability of scores on measure x.

τx: a true score for test x.

e_x: the error portion of an observed score on test x.

β_x: the population regression coefficient for variable τx in a regression model relating true scores.

βzx: the population regression coefficient for variable τx in a regression model where all variables are standardized to have a mean of 0 and standard deviation of 1.

β^*_x: the population regression coefficient for variable x in a regression model relating observed scores.

β̂_x: an estimate of β_x based on a sample.

ε: stochastic error term in a regression model.
Acknowledgements

First and foremost I wish to acknowledge my supervisor, Dr. Bruno Zumbo, who has guided my development as a researcher, supported my learning at every turn, and become both a friend and a mentor. He is a model of the scholar that I can only hope to become one day. I also would like to thank my committee members, Dr. Anita Hubley, Dr. Kim Schonert-Reichl, and Dr. Amery Wu, for their valuable feedback on my developing work and for their teaching both in and out of the classroom. I am also grateful for the mentorship and collaboration of Dr. Brent Davis, who transformed the way I see teaching and mathematics, and whose guidance indirectly influenced the work in this thesis. My work also benefited greatly from methods courses and conversations with Dr. Nand Kishor. Lastly, I am happy to have shared so much learning and enjoyment, some of which can be found in this thesis, with Oscar Olvera, the other half of the “cohort,” and with Tavinder Ark and Yan Liu, supportive senior members of the Edgeworth Lab. On a personal note, I wish to thank my parents, who quite literally made all of my endeavors possible, my sister, for her unwavering faith in me, my aunt and uncle, who welcomed me to Vancouver and introduced me to a wonderful community of scholars, and Lizzy, for too much to describe.
1 Introduction

1.1 Measurement Error and Hypothesis Testing in Regression

Regression is one of the most commonly used data analytic techniques in educational and psychological research, and its use appears to be increasing (Skidmore & Thompson, 2010). Despite the assumption of ordinary least squares (OLS) regression that the analyzed data contain no measurement error, the data analyzed in education and psychology nearly always contain some amount of random measurement error. These data include, but are not limited to, standardized tests, performance assessments, demographic data, and self-report surveys or questionnaires. The common belief seems to be that, if anything, random measurement error will reduce observed effects, making inferences about true effects conservative. In fact, random measurement error in the independent variables of a multiple regression model can have disastrous consequences by increasing Type I error rates (Brunner & Austin, 2009), rendering the associated hypothesis tests invalid and raising the potential for false positive findings in the research literature. This thesis uses computer simulations to document and explain the problem in a non-technical manner for applied researchers, focusing on the implications for research practice and pedagogy.

1.2 Motivating Examples

Validity is considered the primary concern when developing new tests or evaluating test-based inferences and decisions (American Educational Research Association, American Psychological Association, & National Council on Measurement
Validation involves gathering evidence to support arguments (e.g., Kane, 1992, 2006) or explanations (e.g., Zumbo, 2007, 2009) that justify proposed inferences and decisions based on observed test scores. Recent reviews of validation research suggest that correlational evidence, such as predictive and concurrent criterion-related or incremental validity evidence, is the most common way to support test-based inferences and decisions (Cizek, Rosenberg, & Koons, 2008; Hogan & Agnello, 2004; Zumbo & Shear, 2011). Predictive and concurrent criterion-related validity evidence evaluates how accurately new and existing tests are able to predict criterion or outcome variables of interest either now (concurrent) or in the future (predictive; American Educational Research Association et al., 1999). Incremental validity evidence is used to assess whether these tests can add to predictions and decision-making above and beyond other existing tests or sources of information (Hunsley & Meyer, 2003; Sechrest, 1963). This is considered by some to be the “litmus test for determining the utility of a new measure” (Antonakis & Dietz, 2011, p. 49).

Multiple regression is commonly used to evaluate criterion-related and incremental validity evidence. In some cases, a model predicting the criterion using scores from the test under study will be built, and if the regression coefficient for the test scores is significant, it supports the criterion-related validity of the scores. This is common, for example, with college admissions tests such as the SAT (e.g., Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008), or in personnel psychology when evaluating tests for personnel selection (e.g., Van Iddekinge & Ployhart, 2008). In a recent special issue on incremental validity in the journal Psychological Assessment, multiple authors recommended using hierarchical multiple regression analyses to evaluate incremental
validity (Haynes & Lench, 2003; Hunsley & Meyer, 2003). If the test or measure being studied contributes a statistically significant amount to predicting the criterion when entered in the second stage of a regression model, this is taken as evidence of incremental validity, both when considering use of a test for a new purpose or developing a new test. Both of these methods yield identical results.

Researchers may also use multiple regression to test whether one construct is related to another, after taking into account a third (possibly confounding) construct. For example Meinz and Hambrick (2010) evaluated whether working memory capacity (WMC; considered to be an innate ability) could explain a statistically significant amount of variance in piano-playing ability, after controlling for amount of deliberate practice. Based on the results of a hierarchical regression analysis, the authors concluded that innate abilities are uniquely related to observed performance, above and beyond deliberate practice, suggesting that WMC does indeed have a unique effect on observed performance.

In all of these examples, the independent (predictor) variables used in the multiple regression models are measured with error, and are likely to be correlated with other predictors in the model. Under these conditions, results in statistics show that Type I error rates of the hypothesis tests can be drastically inflated (Brunner & Austin, 2009). From a methodological perspective, the hypothesis tests may no longer be valid and hence can neither be recommended nor relied upon. From a pragmatic perspective, this suggests that many research findings in these areas may in fact be false positive results, meaning that high stakes testing decisions or theoretical understandings are based not on reality, but on statistical artifacts. Statistical artifacts in validation research are particularly problematic,
because incorrect conclusions could be carried forward to subsequent studies that rely upon the test or measure evaluated in the validation study.

1.3 Rationale

Although it is widely recognized that measurement error can lead to biased regression estimates in methodological overviews (e.g., Fleiss & Shrout, 1977; Ree & Carretta, 2006), introductory regression textbooks (Cohen, Cohen, West, & Aiken, 2003; Fox, 2008; Pedhazur, 1982), and advanced methodological literature (Bollen, 1989; Lu, Thomas, & Zumbo, 2005; Muthén, 2002), there is much less discussion about the impacts on associated hypothesis tests. There have been discussions of the potential for estimating spurious partial correlations when variables are measured with error (Brewer, Campbell, & Crano, 1970; Burks, 1926; Kahneman, 1965; Lord, 1974; Strauss, 1981), or for spurious findings in ANCOVA (Culpepper & Aguinis, 2011; Lord, 1960; Zinbarg, Suzuki, Uliaszek, & Lewis, 2010) when the covariate is measured with error. However, despite passing references to the connection between partial correlation and multiple regression (Strauss, 1981, p. 3), or presenting a framework that can be used for regression but focuses on ANCOVA (Zinbarg et al., 2010), the problem has not been made explicit in a general way outside the statistical literature.

Statisticians have tended to focus on developing solutions to the problems caused by measurement error (e.g., Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Fuller, 1987; Jöreskog, 1978; Muthén, 2002) rather than documenting the nature and extent of the problems. A similar phenomenon has occurred in robust statistics, where sophisticated methods have been developed to account for outliers (e.g., Maronna, Martin, & Yohai, 2006), but are not widely applied (Liu & Zumbo, 2012). In order to
avoid incorrect data analytic methods, and to encourage researchers to make use of new statistical techniques, there is a need to demonstrate and document the problems that measurement error can cause in multiple regression.

1.4 Study Purpose

The purpose of this thesis is to demonstrate the problems that random measurement error in the independent variables of a multiple regression model can cause for an audience of applied researchers. To do so, computer simulations will be used to document and explain, in a non-technical manner, the nature and extent of inflated Type I error rates in applied research contexts. Computer simulations are an effective tool in this situation because they provide evidence regarding whether mathematical and statistical results are applicable to realistic scenarios, and because they can be used to illustrate the nature of the problem without reliance on complex mathematical expressions.

1.5 Outline of the Thesis

The organization of the thesis is as follows. Chapter 2 provides a general literature review, covering background concepts, relevant psychometric and statistical literature, and previous results. The problem and purpose of the thesis is then stated more precisely. Chapter 3 presents Study 1, a Monte Carlo simulation demonstrating the impacts of measurement error in simple regression. Chapter 4 presents Study 2, a three-part Monte Carlo simulation exploring and demonstrating the impacts of measurement error on Type I error rates in multiple regression, with an emphasis on situations commonly encountered in practice. Chapter 5 summarizes the findings, discusses implications for practitioners and educators, describes novel contributions of the thesis, and lists some limitations and future directions for research.
2  Background and Literature Review

2.1  Chapter Summary

This chapter reviews the concepts of measurement error and reliability within the framework of classical test theory (CTT), describes the notation used in the thesis, clarifies important terms, reviews relevant literature from psychometrics and statistics, and ends with a precise statement of the research questions and purpose.

2.2  Measurement Error and Reliability

The term “measurement error” is used often in the quantitative and methodological literature, but it is not always well-defined. Put simply, measurement error is the difference between the true value of a construct or quantity that we wish to study and the observed value we obtain when we measure it. In some cases, this is a conceptually simple difference – there may be slight variability in measurements of a person’s height, but we have a good idea of what we mean by their “true” height. In many other cases relevant to education and psychology, defining and conceptualizing the true value is more difficult. If we are measuring knowledge of mathematics, what is a person’s “true” level of mathematical knowledge? Often we can never know this exactly, and so we develop frameworks that allow us to meaningfully use such concepts. Even in this case, however, the concept of measurement error is useful. As Kane (2010) described:

Errors of measurement arise when we adopt a conceptual framework that presumes that the construct being measured is invariant over some conditions of observation. If we interpret our observations in terms of general attributes or constructs of persons that should not vary over certain conditions of observation, and the scores do vary over these conditions of observation, we need errors of measurement to resolve the discrepancies (p. 7).
The error is still the difference between the true value we are interested in studying, or making inferences about, and the one that is observed. The primary focus of this thesis is random measurement error, which will be defined and described from within the framework of classical test theory (CTT; Novick, 1966).

In CTT, the observed scores obtained from a measure are conceptualized as consisting of a true score component and a random error component. The true score is defined as the expected value of a person’s observed score across repeated occasions with the same measure, while the error term represents the deviation from that expected value on any particular testing or measurement occasion. The error term represents random measurement error. When we are considering only a single test or measure, we can write the familiar CTT model:

\[ x_{ij} = \tau_{xi} + e_{xij} \]

where \( x_{ij} \) is the observed score for person \( i \) on occasion \( j \), \( \tau_{xi} \) is the true score for person \( i \) on test \( x \), and \( e_{ij} \) is the measurement error for person \( i \) on occasion \( j \) on test \( x \). Using \( E[\cdot] \) to denote the expected value, because \( E[x_{ij}] = \tau_i \), it follows that \( E[e_{ij}] = 0 \). In the remaining text, the subscripts \( i \) and \( j \) are dropped. Based on additional assumptions in the CTT model, \( Cor(\tau_x, e_x) = 0 \) and \( Cor(e_x, e_u) = 0 \), where \( Cor(\cdot) \) denotes the correlation between random variables and \( e_u \) are the errors arising from taking another test, \( u \). Although not strictly necessary within CTT, we make the further common assumption that the error variance is homogeneous across all levels of the true score. We will also make the assumption that all true scores and errors are normally distributed, which implies that the observed scores are also normally distributed.
The properties above are what lead this to be termed random measurement error, because the measurement errors have a mean (expected value) of 0, they have a constant variance for all people, they are uncorrelated with a person’s true scores, and they are uncorrelated with other measurement errors. From a statistical perspective, this is often referred to as classical measurement error (Carroll et al., 2006, p. 2). More specifically, this measurement error is unbiased (because the expected value of observed scores equals the true scores), additive (errors are added to the true score), and non-differential (the errors are uncorrelated with other true scores and errors).

Reliability, which quantifies the amount of measurement error, is defined as the proportion of observed score variance comprised by true score variance for a particular group. Reliability is used as an indicator of how consistent scores are likely to be across different occasions. Formally, we define the reliability of scores on test $x$ as:

$$rel(x) = \frac{Var(\tau_x)}{Var(\tau_x) + Var(e_x)} = \frac{Var(\tau_x)}{Var(x)}$$

where $Var(x)$ represents the variance of the random variable $x$. Under the various formulations, axioms, and definitions of CTT, it can be shown that this is equivalent to the correlation between observed scores from two parallel versions of the same test, or the square of the correlation between the true scores and observed scores (which can never actually be calculated, because true scores are unknown). Although these relationships have allowed researchers to formulate many methods to quantify and estimate the reliability of test scores for particular groups, here the focus is on the concept of reliability rather than particular methods for estimating reliability. See Crocker and Algina (1986), Haertel (2006), Novick (1966), or Zimmerman and Zumbo (2001) for
more detailed overviews of reliability theory and methods for estimating the reliability of a set of test scores.

There are four commonly discussed sources of random measurement error in educational and psychological contexts: “(1) inherent variability in human performance, (2) variations in the environment within which measurements are obtained, (3) variation introduced by the processes of evaluating responses, and (4) error arising from the selection of test question asked or the behavioral sample observed” (Haertel, 2006, p. 68). Schmidt and Hunter (1999) elaborate on these sources of error and their relevance. Clearly these sources of error will be relevant in nearly all educational and psychological contexts.

These concepts in CTT can also be translated into latent variable models, but such extensions are unnecessary for the purpose of this thesis, and are beyond the scope of this review. See Bollen (1989, pp. 218–219) for an overview and some references to alternative formulations. Perhaps the most important point to note is that the translation is more complicated than simply labeling the true score, \( \tau \), as the latent variable in such models. The remainder of this thesis will focus on the CTT model.

### 2.3 Motivating Example, Notation and Definitions

To provide a concrete example that will guide the demonstration, we build on an example presented by Maxwell and Delaney (1993) in a related study. In their example, Maxwell and Delany consider a case in which a researcher has three variables: “number of errors made in a laboratory cognitive task \( (X_1) \), speed of response during the task \( (X_2) \),
and score on a standardized ability test (Y)” (p. 182). The researcher is interested in determining the effects of number of errors and response time on standardized test performance.

The researcher may in fact be interested in whether there is an effect of spatial ability (a cognitive skill indicated by number of errors) or processing fluency (indicated by response time) on mathematics achievement (indicated by score on the standardized test). In this case, interest is not on the specific observed values above, which include variability due to true abilities and random measurement error, but on the true scores – the value we would expect across other occasions and versions of the tasks. Of course, as Schmidt and Hunter (1996) point out, the specific measurements themselves, such as number of errors or response time, even if taken in a laboratory setting, are likely to contain random measurement errors due to the equipment used or human coding errors.

A slightly revised version of the scenario described by Maxwell and Delaney (1993) is as follows. Let \( \tau_u \) represent each student’s true score on a spatial ability task, \( \tau_x \) be the true score for processing fluency, and \( \tau_y \) be the true score on a mathematics achievement test. Supposing mathematics achievement depends upon spatial ability and processing fluency, we can describe the effects among true scores as:

---

1 Zinbarg et al. (2010) provide a similar situation, but replace \( X_1 \) and \( X_2 \) with anxiety and depression, and let \( Y \) be any dependent variable of interest.

2 The term “effect” will be used to refer to directional relationships (e.g., regression coefficients), whereas the term “relationship” will be used generally to refer to both correlations (non-directional) and regression coefficients (directional). We do not consider issues involved in determining whether a directional or causal inference is warranted, which is an issue of experimental design; here we assume a causal or directional inference is warranted and study how the estimate of that effect is likely to be impacted.

3 This example also benefits from slight revision because Maxwell and Delaney consider a case in which \( X_1 \), the number of errors on a cognitive task is positively correlated to the standardized ability score, when in fact it would likely make more sense if number of errors were negatively correlated with the ability score. Here we use the total score (i.e. number correct or number of non-errors) on the cognitive ability test rather than the number of errors (i.e. number incorrect).
where $\epsilon \sim N(0, \sigma_{\epsilon})$, and represents the stochastic portion of the regression model, which is distinct from measurement error. Note that this scenario is an instance of multiple regression rather than multiple correlation; that is, the independent variables ($\tau_x$ and $\tau_u$) are causally prior to the dependent variable $\tau_y$, and so there is a directionality in the relationships, meaning they are effects.

Unbeknownst to the researcher, however, suppose mathematics achievement does not depend on processing fluency, meaning that $\beta_x = 0$. Equation 1 can then be rewritten as:

$$\tau_y = \beta_0 + \beta_u \tau_u + \epsilon$$

Figure 1 depicts the relationships among the true scores graphically. Note that there is no path going from $\tau_x$ to $\tau_y$ because mathematics achievement does not depend upon processing fluency. Although there may be a correlation (relationship) between $\tau_x$ and $\tau_y$, the figure makes it clear that this will only be a confound due to the fact that processing fluency is related to spatial ability, and there is an effect of spatial ability on mathematics achievement.

![Figure 1. Graphical Depiction of the Relationship Among True Scores.](image)

Because the true scores cannot be observed, the researcher actually gathers data consisting of $u$, $x$, and $y$, which are the observed scores for spatial ability, processing
fluency, and mathematics achievement, respectively. We can represent the relationships among the observed scores as:

\[ y = \beta_0^* + \beta_u^* u + \beta_x^* x + \epsilon^* \]  

(2)

The “*” distinguishes these from the true score coefficients, emphasizing that the models in Equations 1 and 2 may or may not be the same. The primary question of interest is, if researchers are interested in studying the effects of spatial ability and processing fluency on mathematics achievement, but they base their inferences on the \( \beta^* \) coefficients, what are the likely impacts going to be? More specifically, what are they likely to conclude about the effect of processing fluency, given that the true effect is 0? Before turning to the literature to gain insight into this question, some key terms are briefly discussed.

**Bias and Type I Errors**

The conventional “\( \hat{\cdot} \)” notation will be used to indicate a sample-based estimate of a population parameter. For example \( \hat{\beta}_x^* \) represents the estimated regression coefficient for \( x \) based on a sample of observed scores. A sample statistic, such as \( \hat{\beta}_x^* \) is said to be unbiased if its average value equals the true population value, in this case \( \beta_x^* \). In other words, if \( E[\hat{\beta}_x^*] = \beta_x^* \), it is an unbiased estimate. However, because researchers are actually interested in knowing the value of \( \beta_x \) rather than \( \beta_x^* \), we will consider \( \hat{\beta}_x^* \) to be an unbiased estimate only if \( E[\hat{\beta}_x^*] = \beta_x \), even though statistically this is a slightly different use of the term “bias.”

A similar distinction will be used to operationalize a Type I error. Statistically, a Type I error occurs when we reject a null hypothesis that is in fact true. In the example here, the null hypothesis \( H_0: \beta_x = 0 \) is true, as shown in Equation 1. However, the null hypothesis for Equation 2, \( H_0^*: \beta_x^* = 0 \) may or may not be true. When a researcher gathers
data (on observed scores) and is interested in testing whether $\beta_x$ is non-zero, the inference will actually be based on $\hat{\beta}_x^*$. There are at least two equivalent hypothesis tests that can be used (Cohen et al., 2003). The first is the common $t$-test of the regression coefficient, which has the form:

$$
    t = \frac{\hat{\beta}_x^*}{SE_{\hat{\beta}_x^*}}
$$

(3)

The other is to test the change in the model $R^2$ that results from adding $x$ into the regression model, which has the form:

$$
    F = \frac{\Delta R_{x}^{*2}}{1 - R^{*2}} \frac{n - 2 - 1}{1}
$$

where $R^{*2}$ is the estimated $R^2$ from regressing $y$ on $u$ and $x$ (Equation 2 above), and $\Delta R_{x}^{*2}$ is the increment to that $R^2$ gained by including $x$ in the model. Statistically, these two tests will yield identical results.

Technically, both of these statistics test the null hypothesis $H_0^*: \beta_x^* = 0$, even though researchers are actually interested in testing $H_0: \beta_x = 0$. To see the difference, consider the case where $H_0$ is true (as it is above), but $H_0^*$ is false, meaning that $\beta_x^* \neq 0$. We conduct the $t$ (or the $F$) test above and the result is statistically significant; in this case we would (correctly) reject $H_0^*$. However, we have incorrectly rejected $H_0: \beta_x = 0$, which is the hypothesis that researchers are actually interested in. We define this rejection as a Type I error because $H_0$, the “conceptual null hypothesis” (DeMars, 2010, p. 963), the one researchers actually care about, is in fact true.

To summarize, bias and Type I errors will be defined in relation to the underlying true score effect, $\beta_x$. As DeMars (2010) points out, if researchers were “literally only interested in one particular test form on one particular testing occasion” (p. 963), then it
would make sense to use $\beta_1^*$ as the reference point. However, since researchers are interested in the underlying proficiency, and generalizing beyond a single moment in time, conceptually it makes more sense to use $\beta_x$ as the reference point. This distinction will be further elaborated in Chapter 5.

**Standardized vs. Unstandardized Coefficients**

To facilitate interpretation in cases where variables may not have an inherently meaningful scale, researchers in education and psychology frequently rely on standardized regression coefficients (Cohen et al., 2003). Here, $\beta_{zx}$ or $\beta_{zx}^*$ is used to denote a standardized regression coefficient, which is the regression coefficient when all scores, including the criterion, are standardized to have a mean of 0 and standard deviation of 1. Standardized and unstandardized regression coefficients are related by the following formulas:

$$
\beta_{zx} = \beta_x \frac{\sqrt{\text{Var}(x)}}{\sqrt{\text{Var}(y)}} \\
\beta_x = \beta_{zx} \frac{\sqrt{\text{Var}(y)}}{\sqrt{\text{Var}(x)}}
$$

### 2.4 Review of Psychometric and Statistical Literature

#### 2.4.1 Attenuation of Correlations

It has long been recognized that when random measurement error is present, as in the CTT model described above, the correlation coefficient between two observed variables will be biased towards zero (e.g. Cohen et al., 2003; Crocker & Algina, 1986; Spearman, 1904, 1910). This phenomenon is referred to as “attenuation” (Spearman, 1904, p. 89), and is closely related to many fundamental results in CTT. In the scenario
from the prior section, the correlation between true scores on the processing fluency task mathematics achievement tests can be related to the observed score correlation by the following:

$$ \text{Cor}(y, x) = \text{Cor}(\tau_y, \tau_x) \sqrt{\text{rel}(y) \times \text{rel}(x)} $$

Because reliabilities are always less than or equal to 1, the observed score correlation will always be less than (i.e., closer to zero) or equal to the true score correlation. Hence, when estimating the correlation between two true scores based on a correlation between observed scores, researchers often use a “corrected” or disattenuated correlation, which is:

$$ \text{Cor}_{adj}(y, x) = \frac{\text{Cor}(y, x)}{\sqrt{\text{rel}(y) \times \text{rel}(x)}} $$

where $\text{Cor}_{adj}(y, x)$ is the estimated adjusted or disattenuated correlation. Here errors are assumed to be uncorrelated with one another and with the true scores.

In the case of correlations, the effect of measurement error is the same regardless of which variable contains error ($y$ or $x$). Moreover the bias will always be towards 0, so that observed relationships are always weaker than the true relationship. In the case where $\text{Cor}(\tau_y, \tau_x) = 0$, however, there is no bias because $\text{Cor}(y, x) = \text{Cor}(\tau_y, \tau_x) \times 0$ is also 0. Hypothesis tests using $\text{Cor}_{adj}(y, x)$ are complex, and there is disagreement about the best way to do conduct these tests (Charles, 2005; Raju & Brand, 2003; Schmidt & Hunter, 1999), but the Type I error rate of the test of the unadjusted (attenuated) correlation will not be inflated.

2.4.2 Bias and Type I Error Rates in Simple Regression

Instead of the correlation between two variables, one may be interested in the
regression coefficient of an independent variable, $x$, for a dependent variable, $y$. The impacts of measurement error in this case, referred to as simple regression, are well documented (e.g. Carroll et al., 2006; Cohen et al., 2003; Gustafson, 2004). There are four possible scenarios to consider: no measurement error, error in $y$, error in $x$, or error in both $x$ and $y$. We can also distinguish between the impacts on standardized versus unstandardized regression coefficients. The standardized regression coefficient in simple regression is equivalent to the correlation between $x$ and $y$; hence measurement error in either $x$ or $y$ will attenuate the standardized regression coefficient, a result that follows directly from the previous section.

When regression coefficients are unstandardized (in the same units as observed scores), measurement error in $y$ will not bias the estimated regression coefficient (Cohen et al., 2003, pp. 56-57). However, measurement error in $y$ will impact the estimate of the residual variance ($Var(\varepsilon)$), which will impact the standard error of prediction and the estimated standard error of the regression coefficients (Gustafson, 2004; Pedhazur, 1982). Meanwhile, measurement error in $x$ will attenuate the estimated coefficient, biasing it towards 0. Specifically, $\beta_x^* = rel(x)\beta_x$. Again, because $rel(x) < 1$ (except where there is no measurement error), the coefficient based on observed scores will always be smaller than the true regression coefficient. Overall then, bias in simple regression will either make observed coefficients closer to zero, or will have no impact, depending upon which variable(s) has error and which type of coefficient we estimate.

When the null hypothesis is true and $\beta_x = 0$, there will be no bias in standardized or unstandardized regression coefficients, because $\beta_x = 0$ implies $\beta_x^* = rel(x) * 0$ is also zero, as in the case of correlation. In fact, the $t$-statistic for the estimated coefficient will
be distributed exactly as a \( t \)-distribution with \( n-2 \) degrees of freedom, and the Type I error rate of the test will be exactly correct (Buonaccorsi, 2010, p. 81; Carroll et al., 2006, p. 244). However, an inspection of the formula for the standard error,

\[
\hat{SE}_{\hat{\beta}_x} = \sqrt{\frac{\hat{\sigma}^2_e}{\sum (x_i - \bar{x})^2}}
\]

suggests an interesting phenomenon. Specifically, measurement error in \( y \) will inflate the estimated residual variance, which will in turn inflate the estimated standard error. In addition, adding measurement error to \( x \) will tend to inflate its variance and hence deflate the estimated standard error. The fact that the \( t \)-test based on this estimated standard error remains exactly correct suggests a complex interplay of factors, something that will be demonstrated in Study 1 in Chapter 3 of this thesis. Note that, although the \( t \)-test under the null is correct (hence maintaining correct Type I error rates), the test is not necessarily correct or optimal when the null hypothesis is false (\( \beta_x \neq 0 \)).

### 2.4.3 Bias in Multiple Regression

In the case of multiple regression, the extent and nature of the bias can be hard to describe or predict, even for the simplest measurement error scenarios, and has been described as “complicated” on more than one occasion (Carroll et al., 2006, p. 52; Cochran, 1968, p. 655). To provide an initial sense for the impacts of measurement error in multiple regression, we return to the scenario above, relating spatial ability, processing fluency, and mathematics achievement. This discussion of the two-predictor regression case follows extensive treatments in the statistical literature (e.g., Carroll et al., 2006; Cochran, 1968; Gustafson, 2004), but focuses specifically on the case outlined above, in which \( \beta_x = 0 \).
Suppose that a researcher gathers data on spatial ability, processing fluency, and mathematics achievement to test the effect of processing fluency on achievement, after taking into account spatial ability. As in the scenario above, there is no true effect of processing fluency ($\beta_x = 0$), but the researcher will base his or her inferences on the observed scores $x$, $u$, and $y$. Following Maxwell and Delaney (1993), suppose that for the model in Equation 1 and Figure 1, $\beta_u = 10$, $Var(\tau_u) = 1$, $Var(\varepsilon) = 104$, $Cor(\tau_x, \tau_u) = 0.5$, and $\beta_0$ is zero. This implies that $Cor(\tau_y, \tau_u) = 0.7$ and $Cor(\tau_y, \tau_x) = 0.35$. We also assume that $Var(\tau_x) = 1$, and that the means of $\tau_u$ and $\tau_x$ are 0. This implies that $Var(\tau_y) = 204$, that the mean of $\tau_y$ is zero, and that the standardized coefficient, $\beta_{zu} = 0.7$. In addition, suppose that the reliabilities of $x$, $u$ and $y$ are $rel(x) = rel(u) = rel(y) = 0.8$. Assuming that the true score model remains the same, and the random measurement error is added as noise, the mean of all observed scores remains zero, while the variances of the observed scores are $Var(x) = Var(u) = 1.25$ and $Var(y) = 255$.

We will use the following expression for estimating standardized regression coefficients in terms of correlations (Cohen et al., 2003) to understand how measurement error will bias the observed estimates:

$$
\beta_{zx}^* = \frac{Cor(y, x) - Cor(y, u)Cor(x, u)}{1 - Cor(x, u)^2}
$$

---

4 Maxwell and Delaney (1993) assume an intercept of $\beta_0 = 100$, which may be a reasonable value when using standardized achievement test scores. However, as we will see, the estimated effects do not depend in any way on the variable means, and so this assumption implies no loss of generality of the results.

5 $Var(\tau_y) = Var(10 * \tau_u) + Var(\varepsilon) = 100 * Var(\tau_u) + 104 = 204$

6 In these examples, and below, only cases where $Cor(\tau_x, \tau_u) \geq 0$ are considered. If $\beta_x = 0$, we can see that when $Cor(\tau_x, \tau_u) < 0$, there will still be bias and the estimated coefficients will just be multiplied by negative 1. We limit the discussion to positive correlations because a) conclusions about the nature and extent of the problem will generalize to negative correlations and b) we are more often interested in testing for a confound when two variables are positively correlated.
\[ \beta_{zu}^* = \frac{\text{Cor}(y, u) - \text{Cor}(y, x)\text{Cor}(x, u)}{1 - \text{Cor}(x, u)^2} \]

Note that these are all observed score correlations. Using true score correlations provided above yields the correct estimates of the various parameters (e.g., \( \beta_{zu} = 0.7 \) and \( \beta_{zx} = 0 \)). However, using the observed (attenuated) correlations, based on a reliability of 0.80, yields an estimate of \( \beta_{zu}^* = 0.067 \). Using the formula to convert this to an unstandardized coefficient we get \( \beta_x^* = 0.957 \). Meanwhile, similar formulas yield estimates of \( \beta_{zu}^* = 0.533 \) and \( \beta_u^* = 7.617 \). Measurement error has yielded estimates of \( \beta_u^* \) that are too small and estimates of \( \beta_x^* \) that are too large. A particularly problematic fact is that while \( \beta_x \) is in fact zero, our estimated effect, \( \beta_x^* \), is non-zero, meaning that the estimated effect is spurious – a statistical artifact of measurement error. We would say that these estimates are biased. And, unlike correlation or simple regression, where biases were always downwards, and when the true value is zero there was no bias, in multiple regression the bias can either be upwards or downwards, or even yield an estimated effect where there is no true effect.

Note that if there is error in \( y \) only, we have the result that:

\[ \beta_{zx}^* = \beta_{zx}\sqrt{\text{rel}(y)} \]

and

\[ \beta_x^* = \beta_{zx}^* \sqrt{\frac{\text{Var}(y)}{\text{Var}(\tau_y)}} = \left( \beta_{zx} \frac{\sqrt{\text{Var}(\tau_y)}}{\sqrt{\text{Var}(y)}} \right) \sqrt{\frac{\text{Var}(y)}{\text{Var}(\tau_x)}} = \beta_{zx} \frac{\sqrt{\text{Var}(\tau_y)}}{\sqrt{\text{Var}(\tau_x)}} = \beta_x \]

so that measurement error in \( y \) only will attenuate the standardized coefficients, but it will not bias the unstandardized coefficients. Similar expressions can also be used to show that the bias due to measurement error in \( y \) will be the same whether there is
measurement error in \( x \) and \( u \) or not. Although the impacts of measurement error in \( y \) are less often discussed (Buonaccorsi, 2010), this difference has important implications for applied researchers, something discussed further in Chapter 5.

Because measurement error in \( y \) will not bias unstandardized coefficient estimates beyond the error in \( x \) and \( u \), we can use the following result from Gustafson (2004, p. 19, Equation 2.8, modified to fit present circumstances), which assumes that \( \tau_x \) and \( \tau_u \) are standard normal variables, measurement errors behave as described above for the CTT model, and there is no measurement error in \( y \):

\[
\beta_x^* = \frac{(1 + \text{Var}(e_u) - \text{Cor}(\tau_x, \tau_u)^2)\beta_x + (\text{Cor}(\tau_x, \tau_u)\text{Var}(e_u))\beta_u}{(1 + \text{Var}(e_x))(1 + \text{Var}(\tau_u)) - (\text{Cor}(\tau_x, \tau_u)^2)}
\]

A similar, symmetric equation, is true for \( \beta_u^* \). This expression is quite complicated, but it can provide some insight into the case where \( \beta_x = 0 \). If \( \beta_x = 0 \), then \( \beta_x^* \) will also be 0 if there is no measurement error in \( u \), or if \( \tau_x \) and \( \tau_u \) are uncorrelated. Looking at the formulas above, we can tell that when the true predictors are uncorrelated, the multiple regression coefficients actually reduce to the simple regression coefficients and the bias follows the same results discussed above. Note also that even if there is no measurement error in \( x \), there could still easily be a non-zero estimate of \( \beta_x^* \).

To summarize, if we are interested in estimating the effect of \( x \) in the observed score regression model \( (\beta_x^*) \) when the true value of \( \beta_x \) is 0 but \( \beta_u \neq 0 \), then we can summarize some relevant results from the above sources as follows:

- If there is no error in \( u \), then estimates of \( \beta_x^* \) will be correct (i.e., they will be 0).
- If \( \text{Cor}(\tau_x, \tau_u) = 0 \), then estimates of \( \beta_x^* \) will also be correct, even if there is measurement error in \( u \).

Note that while some of these results generalize to all cases, some are only true for the case here, where \( \beta_x = 0 \).
• Measurement error in $y$ only will not bias estimates of $\beta^*_x$ when $\beta_x = 0$.

• When $\text{Cor}(\tau_x, \tau_y) \neq 0$ and there is error in $x$ only, estimates of $\beta^*_x$ will not be biased but estimates of $\beta^*_u$ could be.

• When $\text{Cor}(\tau_x, \tau_y) \neq 0$ and there is error in $u$ only, both estimates of $\beta^*_x$ and $\beta^*_u$ can be biased; estimates of $\beta^*_x$ will be too large (i.e., non-zero) and estimates of $\beta^*_u$ will generally be too small.

• When $\text{Cor}(\tau_x, \tau_u) \neq 0$ and both predictors are measured with error, the resulting biases are complex to describe and will depend upon the correlation between the predictors, the variance of the measurement error (magnitude of reliabilities), and the true score parameter values (both regression coefficients and $\text{Var}(\varepsilon)$). In general, all estimates will likely be biased.

It is clear that measurement error can severely bias the estimated regression coefficients, and the nature of the bias (large, small, positive, or negative) will depend upon the correlations among the true scores as well as the variances and covariances of the errors. This makes it difficult to predict the exact impacts in advance (e.g. Levi, 1973). Further, as multiple authors point out (e.g. Bollen, 1989; Buonaccorsi, 2010; Carroll et al., 2006; Gustafson, 2004; Levi, 1973), even if there are multiple predictors and only one is measured with error, all regression effects could still be biased. Perhaps most importantly, and returning to our example, it appears that researchers could conclude that there is an effect between two variables, when in fact there is not. In educational and psychological research, this inference will often depend not only on the precise parameter estimate, but also on the result of the statistical significance test. It turns out that the hypothesis test can lead to the incorrect conclusion as well.

2.4.4 Type I Error Rates in Multiple Regression

Even in the statistical literature, Brunner and Austin (2009, p. 35) note only “passing references – for example by Fuller (1987, p. 55), Cochran (1968, p. 653), Carroll et al. (2006, pp. 52-53) – to incorrect Type I error rates.” Furthermore, none of these references systematically documents the problem. To this list we can add Gustafson
(2004), who notes, “an additional troubling aspect of measurement error is its ability to induce 'apparent significance' in other precisely measured predictors which actually have no effect on the response” (p. 13). Elsewhere, Austin and Brunner (2004) have written that, “although there is a connection between biased estimation and increased type I error rate, all the discussions of residual confounding that we have seen leave the connection implicit” (2004, p. 1160). They have conducted a series of studies (Austin & Brunner, 2003, 2004; Brunner & Austin, 2009) systematically and explicitly documenting various ways that measurement error inflates Type I error rates in regression contexts. These are described below.

Two other recent studies (Culpepper & Aguinis, 2011; Zinbarg et al., 2010) have raised the issue of inflated Type I error rates in analysis of covariance (ANCOVA), something pointed out previously in the statistical (Carroll et al., 2006) and psychometric literature (Lord, 1960). Although not always recognized, ANCOVA is a special type of multiple regression model, where \( \tau_x \) is a (usually dichotomous) grouping variable, and \( \tau_u \) is a continuous covariate we wish to control for. If there are group differences on the covariate measure, this translates to a correlation between \( \tau_x \) (true group membership) and \( \tau_u \). From the results above, we know this can lead to a spurious estimate of group differences on the outcome variable \( y \), even if there are no true differences after controlling for \( \tau_u \).

Zinbarg et al. (2010) combine structural equation modeling and multiple correlation frameworks to show that parameter estimates and Type I error rates will be biased in ANCOVA and analysis of partial variance (APV), when estimates of relationships among latent variables are based on observed variables. The latent variables
in their framework are actually akin to the true scores described above. After providing a formula for the estimated standardized regression coefficient of \( y \) on \( x \) (\( \beta_{zy}^* \) using the notation here) when the true coefficient is 0 (\( \beta_x = 0 \)), they present a table illustrating the bias across seven different examples (Table 1, p. 311). Focusing primarily on the case where \( x \) is a dichotomous grouping variable, they present the bias in the estimated zero-order correlation, standardized regression coefficient, semi-partial correlation, Cohen’s \( f^2 \) effect size, and Cohen’s \( d \) effect size for seven different combinations of the reliability of \( u \), reliability of \( x \), reliability of \( y \), correlation between \( x \) and \( u \), and the correlation between \( u \) and \( y \). They also report the anticipated Type I error rate across five sample sizes (\( N = 60, 80, 100, 200, \) and \( 400 \)) when the nominal Type I error rate is \( \alpha = 0.05 \). They do not provide details regarding how anticipated Type I error rates were calculated, nor is there a systematic investigation covering a wide range of conditions.

Culpepper and Aguinis (2011) derive expressions representing the anticipated bias in the \( \Delta R^2_x \) for the special case of ANCOVA (with \( x \) as the grouping variable), as well as the anticipated Type I error rate of this test when the null hypothesis is true (i.e., \( \Delta R^2_x = 0 \)). Both expressions are based on means, standard deviations, correlations, reliabilities and relative group sizes, and hold for very large sample sizes. Although the primary focus of the study is to compare four suggested methods for correcting this bias, they also include figures demonstrating the anticipated Type I error rate for ANCOVA across a variety of conditions and a sample size of 500, using the analytic derivations presented in their appendix. These conditions vary according to reliability of \( u \), correlation between \( u \) and \( y \), and mean differences in \( u \) across treatment groups.
As Zimmerman (2011) demonstrates, relationships and properties that hold in populations or very large samples may not always be obtained in smaller samples. The results in these two studies, as well as those described in the previous section, based on properties and relationships that hold in the population (e.g., reliability of scores or mean differences), may or may not accurately represent patterns that would occur across small samples. For example, an anticipated Type I error rate that is based on the reliability of \( u \) in the population does not account for sampling variability in the reliability of \( u \) in smaller samples, and may or may not be an accurate representation of the likely Type I error rates across small samples.

In addition to the analytic results for large samples that are generally discussed in the statistical references above, Dunlap and Kemery (1988) were perhaps the first to document the problem of inflated Type I errors in multiple regression in an empirical simulation study. They reported an “unexpected” (1988, p. 256) increase in Type I errors when testing for certain forms of interaction effects in a multiple regression model, but this was not a primary aspect of the study and received little interpretation.

The most systematic discussion of how measurement error impacts Type I error rates in multiple regression, and perhaps the only one to make the problem very explicit for a regression context, is the previously mentioned study by Brunner and Austin (2009). They derive the predicted impact mathematically using theoretical results for very large sample sizes, and then confirm its relevance to finite samples with a simulation study. Using similar results to those above, Brunner and Austin demonstrate mathematically that in the presence of measurement error, the estimate of \( \beta_x \) will often be non-zero even when \( \beta_x \) is zero. Then, using the common \( t \)-test for the regression coefficient presented in
Equation 3, they note that because the standard error in that equation has $\sqrt{n}$ in its denominator, the standard error will tend towards zero as the sample size increases to infinity. As Brunner and Austin describe it: “Consequently, the denominator [of the $t$-statistic] converges to zero, the absolute value of the $t$-statistic tends to infinity, and the associated $P$-value converges almost surely to zero. That is, we almost surely commit a Type I error” (2009, p. 37). In other words, the null hypothesis $H_0: \beta_x = 0$ will almost surely be rejected even though in fact $H_0: \beta_x = 0$ is true. Note that they have used “Type I error” in the same way described above. These results are only true for asymptotically large sample sizes, however.

To evaluate the applicability of these results to finite samples, Brunner and Austin (2009) used a Monte Carlo simulation to study five factors that, mathematically speaking, should impact the Type I error probabilities: 1) reliability of $x$, 2) reliability of $u$, 3) correlation between $\tau_x$ and $\tau_u$, 4) true model $R^2$, and 5) sample size. In the limited results published in the paper, they show that, even when the reliability of each predictor is 0.9 (considered to be very high in educational and psychological research), the Type I error rate can actually reach 1.00. They summarize the findings as:

…we see that except when the latent independent variables $[\tau_x]$ and $[\tau_u]$ are uncorrelated, applying ordinary least squares regression to the corresponding observable variables [x] and [u] results in a substantial inflation of the Type I error rate. As one would predict from [analytic results] with [uncorrelated measurement errors], the problem becomes more severe as $[\tau_x]$ and $[\tau_u]$ become more strongly related, as $[\tau_u]$ and $[\tau_y]$ become more strongly related, and as the sample size increases. We view the Type I error rates…as shockingly high, even for fairly moderate sample sizes and modest relationships among variables…In addition, the Type I error rates increase with decreasing reliability of [u], and decrease with decreasing reliability of [x] (the variable being tested) (Brunner & Austin, 2009, pp. 39-40; variable names are edited to be consistent with the notation used here).

Although not presented (or discussed) in the paper, Brunner and Austin provide the complete results (across 7,500 conditions) on-line, and note that the findings are
consistent across all combinations of reliability studied, and across various forms of the measurement error distribution. In other words, the Type I error rates tend to be most inflated in those conditions where the bias in the estimated coefficient is greater. Framed differently, these results suggest that study characteristics we often strive for – stronger studied effects ($R^2$), larger sample sizes, and greater reliability of $x$ (the effect we are focusing on) – will actually lead to worse problems if there is even minor error in $u$ and some correlation between predictors. Beyond the quotation above, there is no analysis of which factors seem to be most important in driving the problem of Type I error rates, or how the factors might interact in causing inflated Type I error rates.

In addition to being presented in a technical manner geared towards statisticians, the study by Brunner and Austin (2009) leaves important questions for applied researchers. First, they do not study the impact of measurement error in $y$, which will often be encountered in educational and psychological research. Although the literature reviewed above suggests error in $y$ will often not bias estimated coefficients, it may impact other parameters, such as standard errors, which could impact Type I error rates. Second, many of the conditions studied may not be representative of those encountered in educational and psychological research: three of the five correlations between $\tau_x$ and $\tau_u$ they studied were greater than or equal to 0.75 (often considered problematic due to multicollinearity), and the $R^2$ values studied were 0.25, 0.50, and 0.75 – values much higher than those likely to be encountered in practice. For example, an $R^2$ of 0.25 (the smallest one studied) corresponds to an effect size of $f^2 = 0.33$, considered to be nearly a “large” effect size by Cohen (1992). Third, Brunner and Austin did not report (or
record) the impact on additional parameter estimates of interest, such as standard errors and effect sizes.

### 2.5 An Unrecognized Problem

Despite clear problems, researchers routinely draw inferences about relationships and effects among true scores based on the results of data analyses performed on observed scores. In light of recent calls to decrease reliance on significance testing and instead report effect sizes or confidence intervals (Cohen et al., 2003; e.g. Kirk, 1996; Moss et al., 2006), some may find the problem of Type I errors in hypothesis testing irrelevant. Yet as Baugh (2002) reminds us, measurement error can also adversely impact effect sizes. In the case of regression, for example, effect sizes such as $\Delta R^2$ are directly linked to bias in the estimated effects and to the significance of the coefficients, as mentioned above. Perhaps more importantly, however, even among those who recognize the problems of measurement error, there seems to be a widespread belief that measurement error will always produce downward bias.

Looking through widely used regression textbooks (Cohen et al., 2003; Fox, 2008; Pedhazur, 1982) and discussions about how measurement error can impact statistical analyses (Fleiss & Shrout, 1977; Liu & Salvendy, 2009; Osborne & Waters, 2002; Ree & Carretta, 2006), there is reference to the potential for bias, but not inflated Type I error rates. Introductory psychometrics textbooks (Allen & Yen, 1979; Crocker & Algina, 1986; Furr & Bacharach, 2008) discuss attenuation in correlation, but do not cover bias and inflated Type I error rates for multiple regression. References to the problem of spurious significance for ANCOVA dates back to Lord (1960), in addition to those reviewed above, but the connection to multiple regression is not made explicitly.
Numerous papers (and textbooks) explain the potential for spurious partial correlations (Brewer et al., 1970; Burks, 1926; Cohen et al., 2003; Kahneman, 1965; Linn & Werts, 1973; Lord, 1974; Sechrest, 1963; Strauss, 1981). But the connection between partial correlations and multiple regression is only mentioned in passing (if at all); more importantly, researchers may feel that concerns about correlations do not apply to the $t$-tests of regression coefficients or the $F$-tests of incremental $R^2$ tests. There is a need to document how problematic measurement error can be for the intended inferences researchers make and that, in some cases, random measurement error will create spurious relationships rather than masking true relationships.

### 2.6 Purpose of the Study

The purpose of this study is to provide a demonstration that random measurement error can severely distort the results of a multiple regression analysis and lead to highly inflated Type I error rates. Towards that end, this thesis presents a series of Monte Carlo computer simulations across two studies, which will be used to explain and understand how random measurement error can cause these problems in commonly encountered scenarios. These simulations will address the following research questions:

a) Does simple regression have correct Type I error rates in the presence of measurement error and what happens to the standard errors?

b) How does measurement error in $y$ impact the Type I error rates in multiple regression?

c) Of the relevant factors, which are most important in causing Type I error rates to be inflated in multiple regression? Which factors lead to the greatest inflation of Type I error rates?

d) What level of Type I error inflation could be anticipated in a realistic data analysis scenario? How are measures of effect size impacted in this case?

The discussion following the simulations will focus on the implications of the problem for both researchers and educators. The next chapter presents Study 1, a demonstration of
the impact of measurement error in simple regression that addresses question a) above. The subsequent chapter presents Study 2, documenting the impact of measurement error in multiple regression, with a particular emphasis on the sort of cases likely to be encountered in practice, addressing questions b) – d).
3 Study 1: Reviewing and Demonstrating the Impact of Measurement Error in Simple Regression

3.1 Introduction

This study used a Monte Carlo simulation to review and demonstrate how measurement error can impact hypothesis tests in simple regression. Mathematical results show that the test of the null hypothesis should be correct and hence Type I error rates should be maintained at their nominal alpha levels (Buonaccorsi, 2010; Carroll et al., 2006). This simulation verified that result across a range of sample sizes and reliabilities. It also demonstrated that although the resulting t-test is valid, parts of the ratio used to calculate the t-test, such as the estimated standard error, are distorted. This chapter presents the methods and results for the study; a discussion of the results is presented in Chapter 5 of the thesis.

3.2 Methods

3.2.1 Study Design

A Monte Carlo simulation was used to examine the Type I error rates and other relevant statistics (described below) across four different levels of reliability for the dependent variable, \( y \), and the independent variable, \( x \) (0.4, 0.6, 0.8, and 1.00), and four different sample sizes (25, 50, 100, and 500). Because the reliability of both \( y \) and \( x \) were varied, the design was a three-factor experimental design, with \( 4 \times 4 \times 4 = 64 \) conditions. We conducted 1000 replications for each condition.
3.2.2 Data Generation

In each of the 64 conditions, data were generated according to the following scenario for the underlying true scores:

\[ \tau_y = \beta_0 + \beta_x \tau_x + \epsilon \]

where \( \beta_0 = \beta_x = 0 \), and hence the null hypothesis \( H_0: \beta_x = 0 \) is true. Here \( \tau_x \) and \( \epsilon \) were both standard normal random variables, and hence \( \tau_y \) was also a standard normal random variable. Because in practice \( \tau_y \) and \( \tau_x \) are measured with error, in practice one bases estimates and inferences on the observed scores \( y \) and \( x \):

\[ y = \tau_y + e_y \]
\[ x = \tau_x + e_x \]

Where \( e_y \) and \( e_x \) are random measurement errors, uncorrelated with each other or with the true scores, and with mean 0 and variance:

\[ \text{var}(e_y) = \frac{\text{var}(\tau_y)}{\text{rel}(y)} - \text{var}(\tau_y) \quad \text{and} \quad \text{var}(e_x) = \frac{\text{var}(\tau_x)}{\text{rel}(x)} - \text{var}(\tau_x) \]

In each experimental condition, 1,000 samples were generated by first generating two independent vectors of \( n \) standard normal random variables, where \( n \) is the sample size for the condition. These are the true scores \( \tau_x \) and \( \tau_y \). Two independent vectors of \( n \) normally distributed random variables each with mean 0 and variance \( \text{var}(e_x) \) and \( \text{var}(e_y) \) are then generated, which are the measurement errors for \( x \) and \( y \), respectively. To create the observed variables (with measurement error), the first two vectors from each pair were added element-wise to create the observed \( x \) variables, and the second vector from each pair was added element-wise to generate the observed \( y \) scores. Because all variables were uncorrelated, this represents a model generated under the null hypothesis described above. In addition, the measurement errors all represent non-
differential, unbiased, additive measurement error since they are independent of one another and of the true scores. All data were generated and analyzed with the R statistical software package (R Development Core Team, 2012).

3.2.3 Outcome Variables

For each dataset of the 1,000 random samples in each of the 64 conditions, a regression model of \( y \) on \( x \), using observed scores, was estimated. The estimated model in each sample was:

\[
y_i = \hat{\beta}_0^* + \hat{\beta}_x^* x + \varepsilon^*
\]

To test the null hypothesis that the relationship between true scores was 0 (i.e. \( H_0: \beta_x = 0 \)), the standard \( t \)-statistic formed by \( \hat{\beta}_x^*/SE_{\hat{\beta}_x^*} \) was calculated and compared to the critical value of a \( t \)-distribution with \( n - 2 \) degrees of freedom and a significance level of \( \alpha = 0.05 \), two-tailed. The proportion of samples (out of 1000) for which the null hypothesis was rejected represents the observed Type I error rate. The mean estimated regression coefficient \( \hat{\beta}_x^* \), estimated standardized coefficient, \( \hat{\beta}_{zx}^* \), estimated standard error, \( SE_{\hat{\beta}_x^*} \), and estimated \( R^2 \) was also computed across the 1,000 samples.

3.2.4 Data Analysis

Due to sampling error, the observed Type I error rate in any given experimental condition is not expected to be exactly 0.05, even for conditions in which the Type I error rate is theoretically expected to be 0.05. Bradley (1978) suggested that an observed Type

---

8 There are different ways to manipulate the reliability of a score. Here we manipulated the reliability by holding the true score variance constant and increasing (or decreasing) measurement error. This method has the benefit that the true score model and scenario is invariant across conditions, so that changes in outcome variables represent changes due to measurement error but not due to changes in the underlying true scores. A more extensive discussion of this choice and alternatives is presented in the final chapter, under the limitations section.
I error rate below 0.025 or above 0.075 (for a nominal Type I error rate of 0.05) should lead to the conclusion that the hypothesis test was non-robust or invalid under the studied conditions, calling this a “liberal” criterion. His argument rests on the assertion that if Type I error rates could potentially be 50% too large, researchers should not be willing to use the test as valid. This criteria was used to identify experimental conditions in which the Type I error rates were incorrect, and were described as deflated or inflated, respectively; any conditions exceeding the criteria were summarized and described. Descriptive statistics for the other outcome variables were presented in tabular and graphical form.

3.3 Results

Table 1 displays the observed Type I error rates across the 64 experimental conditions. All observed Type I error rates were within the range 0.025-0.075, and hence none of the Type I error rates were considered either inflated or deflated.

---

9 Other options for determining whether a condition is evidence of truly inflated Type I error rates include the construction of confidence regions (e.g., Serlin, 2000) or using hypothesis tests of the null hypothesis \( H_0: \pi = 0.05 \), where \( \pi \) is the true probability of a Type I error. Hypothesis testing yields a similar cutoff for significantly different from 0.05, and Serlin’s approach also requires specifying an (arbitrary) cutoff value to construct intervals. Hence the 0.075 criterion was used for continuity in the literature and for simplicity of interpretation.
Table 1. Type I Error Rates in Simple Regression.

<table>
<thead>
<tr>
<th>rel(x)</th>
<th>rel(y)</th>
<th>Sample Size</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.40</td>
<td>0.40</td>
<td>0.056</td>
<td>0.052</td>
<td>0.04</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.40</td>
<td>0.046</td>
<td>0.062</td>
<td>0.049</td>
<td>0.054</td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>0.40</td>
<td>0.058</td>
<td>0.05</td>
<td>0.051</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>0.40</td>
<td>0.047</td>
<td>0.045</td>
<td>0.052</td>
<td>0.051</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.40</td>
<td>0.04</td>
<td>0.048</td>
<td>0.048</td>
<td>0.055</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.6</td>
<td>0.05</td>
<td>0.044</td>
<td>0.031</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>0.6</td>
<td>0.04</td>
<td>0.047</td>
<td>0.057</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>0.6</td>
<td>0.058</td>
<td>0.051</td>
<td>0.056</td>
<td>0.053</td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>0.40</td>
<td>0.041</td>
<td>0.051</td>
<td>0.043</td>
<td>0.055</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.40</td>
<td>0.052</td>
<td>0.047</td>
<td>0.054</td>
<td>0.054</td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>0.40</td>
<td>0.063</td>
<td>0.038</td>
<td>0.06</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>0.40</td>
<td>0.046</td>
<td>0.042</td>
<td>0.046</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>0.40</td>
<td>0.051</td>
<td>0.044</td>
<td>0.048</td>
<td>0.046</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.8</td>
<td>0.043</td>
<td>0.056</td>
<td>0.038</td>
<td>0.041</td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>0.8</td>
<td>0.058</td>
<td>0.049</td>
<td>0.042</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>0.8</td>
<td>0.044</td>
<td>0.034</td>
<td>0.044</td>
<td>0.042</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 presents the mean and standard deviation of the estimated regression coefficients (unstandardized and standardized), the mean estimated standard error of each coefficient, and the mean estimated residual variance ($\sigma_e^2$) across a subset of the 64 conditions. Note that the standard deviation of the observed coefficient estimates is an empirical standard error for that parameter.

Table 2. Observed Sample Estimates in Simple Regression with Measurement Error.

<table>
<thead>
<tr>
<th>N</th>
<th>rel(x)</th>
<th>rel(y)</th>
<th>$\hat{\beta}_x$</th>
<th>$SE_{\hat{\beta}_x}$</th>
<th>$\hat{\beta}_{xz}$</th>
<th>$SE_{\hat{\beta}_{xz}}$</th>
<th>$\sigma_e^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>100</td>
<td>0.6</td>
<td>0.6</td>
<td>0.007</td>
<td>0.093</td>
<td>0.102</td>
<td>0.007</td>
<td>0.093</td>
</tr>
<tr>
<td>100</td>
<td>0.6</td>
<td>0.8</td>
<td>-0.002</td>
<td>0.087</td>
<td>0.087</td>
<td>-0.002</td>
<td>0.01</td>
</tr>
<tr>
<td>100</td>
<td>0.6</td>
<td>1.0</td>
<td>-0.001</td>
<td>0.08</td>
<td>0.078</td>
<td>0</td>
<td>0.012</td>
</tr>
<tr>
<td>100</td>
<td>0.8</td>
<td>0.6</td>
<td>0.003</td>
<td>0.117</td>
<td>0.117</td>
<td>0.003</td>
<td>0.099</td>
</tr>
<tr>
<td>100</td>
<td>1.0</td>
<td>0.6</td>
<td>0.001</td>
<td>0.129</td>
<td>0.13</td>
<td>0.001</td>
<td>0.1</td>
</tr>
<tr>
<td>100</td>
<td>1.0</td>
<td>1.0</td>
<td>0</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Note: $rel(x)$ = reliability of $x$; $rel(u)$ = reliability of $u$; $N$ = sample size; $\hat{\beta}_x$ = estimated unstandardized regression coefficient; $\hat{\beta}_{xz}$ = estimated standardized coefficient; $SE_{\hat{\beta}_x}$ = estimated standard error for $\hat{\beta}_x$; $SE_{\hat{\beta}_{xz}}$ = estimated standard error for $\hat{\beta}_{xz}$; $\sigma_e^2$ = estimated residual variance for the unstandardized model.
The bottom row of Table 2, where reliability of $x$ and $y$ are both 1.0, represents the ideal case where there is no measurement error (i.e., as if the true score could be observed), and serves as a reference point. Both regression coefficients (standardized and unstandardized) provide unbiased estimates of the population regression coefficients, given that the mean of each coefficient is nearly 0. However, the variability (standard deviation) of the unstandardized estimated coefficient varies as the measurement error does, while the variability of the standardized regression coefficient is stable across all conditions. Error in $x$ deflates variability in the estimated coefficient while error in $y$ inflates variability. For both coefficients, the estimated standard errors are very close to the observed empirical standard errors of the parameter estimates, suggesting that they are a good estimate of the actual sampling variability. The final column shows that measurement error in $y$ tends to inflate the estimate of the residual variance in the regression model, while measurement error in $x$ appears not to affect it.

Figure 2 depicts the estimated standard error of $\beta_x$, the unstandardized coefficient, across all 16 combinations of reliabilities in the study, with a sample size of $n = 100$. Boxplots show the distribution of these estimated standard errors across the 1,000 random samples, while “x” represents the standard deviation of actual estimated coefficients (the empirical standard error) and the “•” shows the mean estimated standard error. The furthest right boxplot, for example, displays the result when there is no measurement error; here the mean and median estimated standard error are roughly 0.10, as was the empirical standard error. In the fourth plot from the right, however, where the reliability of $y$ is 0.4 and the reliability of $x$ is 1.0, the estimated and empirical standard errors are nearly 0.15 across the 1,000 replications. The trend is clear: error in $y$ increases
variability of estimates, error in x decreases variability, and error in both variables tends to cancel itself out.

Figure 3, similar to that in Gustafson (2004) provides an intuitive sense for why this is the case. Panel A of Figure 3 depicts a single sample of $n = 100$ true scores, where the true regression coefficient $\beta_x = 0$. Both true scores are standard normal variables. Also depicted are the estimated regression slopes from 250 samples of $n = 100$ drawn from the same population. Panel B depicts the same sample as Panel A, but with measurement error added to the $x$ scores so that $\text{rel}(y) = 0.6$ but $\text{rel}(x) = 1.0$. Also depicted are the estimated regression slopes from 250 samples of $n = 100$ from the same population, but based on the observed $y$ values containing measurement error. Panel C shows this for $\text{rel}(x) = 0.6$ and $\text{rel}(y) = 1.0$, while Panel D shows this for $\text{rel}(x) = \text{rel}(y) = 0.6$. Because error in $x$ tends to elongate the points horizontally, flatter slopes will be observed, whereas error in $y$ elongates the cloud vertically, so that steeper slopes will be observed.
Figure 2. Estimated Standard Errors in Simple Regression with Measurement Error.

Note: Rel(X) denotes reliability of x; Rel(Y) denotes reliability of y. Results are for 1,000 replications and sample size n = 100.
Figure 3. Variability of Simple Regression Slopes with Measurement Error.

Note: Panel A depicts a sample of $n=100$ true scores ($\tau_y$ and $\tau_x$) with mean 0 and SD 1. Subsequent panels depict the same true scores with and without error added to them; Panel B: $rel(x) = 1.0, rel(y) = 0.6$; Panel C: $rel(x) = 0.6, rel(y) = 1.0$; Panel D: $rel(x) = 0.6, rel(y) = 0.6$. Estimated regression slopes from the specified scenario are shown for 250 samples in each panel.
4 Study 2: False Positives in Multiple Regression Studies when Independent Variables are Measured with Error

4.1 Introduction

This study used a three-part Monte Carlo Simulation to demonstrate how random measurement error can impact Type I error rates in multiple regression across a wide range of conditions. Building on the work of Brunner and Austin (2009) and others (e.g., Cochran, 1968; Gustafson, 2004), this section returns to the two-predictor multiple regression scenario described earlier. We conducted a simulation to study what happens when one or more of the variables is measured with error, but we base our inferences about the relationships among the true scores on the observed scores, without correcting for measurement error. These results will be used to frame a discussion of the problem and to address the following questions:

a) How does measurement error in $y$ impact the Type I error rates in multiple regression?
b) Of the relevant factors, which are most important in causing Type I error rates to be inflated in multiple regression? Which factors lead to the greatest inflation of Type I error rates?
c) What level of Type I error inflation could be anticipated in a realistic data analysis scenario? How are measures of effect size impacted in this case?

Part I studied how six factors (reliability of $x$, $u$, and $y$, correlation between $\tau_\eta$ and $\tau_u$, $R^2$, and sample size) impacted the inflation in Type I error rates across the entire subspace of possible scenarios, addressing a) and b). Part II studied how these factors impact the inflation in Type I error rates across a more restricted range of all possible scenarios, addressing a) and b) more specifically geared towards applied researchers. Part
III estimated the impact of measurement error on Type I error rates and effect sizes for a small set of conditions commonly encountered in practice, addressing c). This chapter presents the methods and results for the study; a discussion of the results is presented in Chapter 5 of the thesis.

4.2 Methods

4.2.1 Study Design

A three-part Monte Carlo simulation was conducted to study the impact of six different factors on the Type I error rate of the $t$-test for the second coefficient in a multiple regression model. The underlying model describing the relationship between the true scores is:

$$\tau_y = \beta_0 + \beta_u \tau_u + \beta_x \tau_x + \varepsilon$$

Where $\beta_0 = \beta_x = 0$ and $\beta_u = 1$. Conceptually, this represents a psychological process in which $\tau_u$ causes $\tau_y$, but $\tau_x$ does not. Hence the null hypothesis, $H_0: \beta_x = 0$, is true. However, true scores cannot be observed and hence inferences are based on the relationships between observed scores, using the following model:

$$y = \beta_0^* + \beta_u^* u + \beta_x^* x + \varepsilon^*$$

where $\beta_0^*$, $\beta_u^*$, and $\beta_x^*$ are the expected value of the coefficients that would be estimated using observed scores measured with error. After fixing the true score regression coefficients to 0 and 1, six factors can be manipulated, with the theoretically possible range for each factor listed in brackets:

1. Reliability of $x$ [$0 – 1$]
2. Reliability of $u$ [$0 – 1$]
3. Reliability of $y$ [$0 – 1$]
4. True model $R^2$ [$0 – 1$]
5. $Cor(\tau_x, \tau_u) [0 - 1]^{10}$
6. Sample size $[10 - 1,000]^{11}$

These factors define a six-dimensional space. Any two-predictor multiple regression based on observed scores will fall somewhere within the space defined by these six factors, with the noted exceptions for sample size and the correlation between the predictors.

In Part I, to study how these six factors impact the Type I error rate of the test $H_0: \beta_x = 0$, 5,000 experimental conditions within the six-dimensional space were generated. For each condition, the reliabilities of $x$, $u$, and $y$, the correlation between $\tau_x$ and $\tau_u$, and the true model $R^2$ were selected by sampling five independent random uniform numbers on the interval 0 to 1. Sample size was selected by sampling a random uniform number, rounded to the nearest integer, on the interval 10 to 1,000. For each of the 5,000 experimental conditions, 1,000 random samples were generated using the values for the experimental condition as population parameters. The Type I error rate was observed across the 1,000 samples.

Although the design of Part I provided insight into how Type I error rates are impacted across the entire range of possible scenarios, many of the experimental conditions may not be representative of research practice. For example, $R^2$ values above 0.75 may be unrealistic, correlations between predictors greater than 0.75 can cause multicollinearity problems, and reliabilities less than 0.60 may not be high enough to support the use of a given measure. In Part II, 2,000 additional experimental conditions

---

$^{10}$ This factor actually has a theoretical range of -1 to +1, but the resulting biases will simply be multiplied by negative 1 in the case where $\beta_x = 0$, so that the amount of bias and Type I error rates will be the same.

$^{11}$ In reality the sample size could be positive integer greater than 0, but we restrict it to this range, which covers a large range of applied research settings.
were generated within the six-dimensional subspace defined by using the following slightly restricted ranges:

1. Reliability of \( x \) [0.60, 1.00]
2. Reliability of \( u \) [0.60, 1.00]
3. Reliability of \( y \) [0.60, 1.00]
4. True Model \( R^2 \) [0.0, 0.75]
5. \( Cor(\tau_x, \tau_u) \) [0.0, 0.75]
6. Sample size [10,1000]

These values were meant to be more representative of the conditions likely to be encountered in practice. Again, for each experimental condition 1,000 random samples were generated using the experimental condition values as population parameters, and the Type I error rate was observed across the 1,000 samples.

Parts I and II provided a sense for how the six factors impacted Type I error rates across large ranges of possible scenarios. However, these studies did not provide a sense for how high Type I error rates might be in specific conditions, or how other observed values, such as estimates of effect size (e.g., Cohen’s \( f^2 \)) or standardized regression coefficient estimates might be impacted. In Part III, the inflation in Type I error rates was observed across a set of fixed levels of each factor. In all conditions, the reliability of \( x \), \( u \), and \( y \) were set at 0.80, the mean estimated reliability found in a recent meta-synthesis of reliability generalization studies that incorporated thousands of research studies (Vacha-Haase & Thompson, 2011). For the other three factors, the following values were used:

1. \( Cor(\tau_x, \tau_u) \): 0.0, 0.1, 0.3, 0.5, 0.6, 0.7
2. \( R^2 \): 0.01961, 0.1304, 0.2539, 0.50
3. \( N \): 50, 100, 500, 1000

This resulted in a completely crossed experimental design with \( 6 \times 4 \times 4 = 96 \) experimental conditions. The values for correlation between \( \tau_x \) and \( \tau_u \) were selected to encompass a
range of values encountered in practice, and to include Cohen’s (1992) suggested categories of “small,” “medium,” and “large” effect sizes for $r$.\textsuperscript{12} The values of $R^2$ were chosen to correspond with Cohen’s suggested guidelines of “small,” “medium,” and “large” values of the effect size $f^2 = \frac{R^2}{1 - R^2}$, as well as one larger value ($f^2 = 1; R^2 = 0.5$). Finally, sample sizes were selected to represent a range of contexts from smaller experimental sample sizes to a larger sample size potentially used in an observational or larger field study.

Overall, 7,096 experimental conditions were studied: 5,000 conditions were generated randomly drawing from the entire theoretically possible space; 2,000 conditions were generated randomly drawing on a more realistic subspace of all possible scenarios; and the final 96 conditions were selected to represent conditions that could be considered typical of applied research. These studies and conditions are summarized below in Table 3.

Table 3. Summary of Study 2 Design Conditions for Parts I-III.

<table>
<thead>
<tr>
<th>Factors and Ranges</th>
<th>Part I</th>
<th>Part II</th>
<th>Part III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability of $x$:</td>
<td>[0.00, 1.00]</td>
<td>[0.60, 1.00]</td>
<td>0.80</td>
</tr>
<tr>
<td>Reliability of $u$:</td>
<td>[0.00, 1.00]</td>
<td>[0.60, 1.00]</td>
<td>0.80</td>
</tr>
<tr>
<td>Reliability of $y$:</td>
<td>[0.00, 1.00]</td>
<td>[0.60, 1.00]</td>
<td>0.80</td>
</tr>
<tr>
<td>True Model $R^2$:</td>
<td>[0.00, 1.00]</td>
<td>[0.00, 0.75]</td>
<td>0.01961, 0.1304, 0.2539, 0.50</td>
</tr>
<tr>
<td>$Cor(\tau_x, \tau_u)$:</td>
<td>[0.00, 1.00]</td>
<td>[0.00, 0.75]</td>
<td>0.0, 0.1, 0.3, 0.5, 0.6, 0.7</td>
</tr>
<tr>
<td>Sample size:</td>
<td>[10,1000]</td>
<td>[10,1000]</td>
<td>50, 100, 500, 1000</td>
</tr>
<tr>
<td>Design</td>
<td>5,000 Random Conditions</td>
<td>2,000 Random Conditions</td>
<td>Full Factorial (96 Conditions)</td>
</tr>
</tbody>
</table>

\textsuperscript{12} Cohen does not distinguish between attenuated and disattenuated correlations. Note these correlation values are the true underlying correlations; the observed correlations will be slightly lower since all variables have reliability = 0.80 in this simulation.
4.2.2 Data Generation

For each of the 7,096 conditions, 1,000 random samples were generated. The approach used by Brunner and Austin (2009), with slight modifications, was used to generate all samples. In all conditions, the values of the experimental factors were treated as population parameters (except sample size).

First, in keeping with the statistical model shown above, and to represent a causal underlying psychological process, \( \tau_x \) and \( \tau_u \) were generated as random variables with a set population correlation, \( \text{Cor}(\tau_x, \tau_u) \), while \( \tau_y \) values were generated as a function of these variables and a random error term.\(^{13}\) The \( \tau_x \) and \( \tau_u \) values were generated as standard normal random variables with a specified correlation using the following process:

\[
\tau_x = a_1 * R_1 + a_2 * R_3 \\
\tau_u = a_1 * R_2 + a_2 * R_3
\]

where \( a_1 = \sqrt{1 - \text{cor}(\tau_x, \tau_u)} \), \( a_2 = \sqrt{\text{cor}(\tau_x, \tau_u)} \), and \( R_i \) are independent standard random normal variables generated with the \text{rnorm()} function in R. Second, \( \tau_y \) was generated as:

\[
\tau_y = \tau_u + \varepsilon
\]

where \( \varepsilon \) is a normally distributed random variable with mean 0 and variance dependent upon the model \( R^2 \) as follows:

\[
\text{var}(\varepsilon) = \frac{\text{var}(\tau_u)}{R^2} - \text{var}(\tau_u)
\]

Next, the observed scores \( x, u, \) and \( y \) were computed as:

\(^{13}\) Note that this random error term is the residual term in the regression model, not a measurement error.
where $e_x$, $e_u$, and $e_y$ are normally distributed independent random variables with mean 0 and variances:

$$\text{var}(e_x) = \frac{\text{var}(\tau_x)}{\text{rel}(\star)} - \text{var}(\tau_x)$$

All simulations were carried out using the R statistical software package (R Development Core Team, 2012), and all random variables were generated using the rnorm() function. Random number seeds were set at the beginning of the process used to generate design conditions and then again at the beginning of the program used to generate all 1,000 samples across the experimental conditions in order to make the results fully reproducible.

4.2.3 Outcome Variables

For each of the 1,000 samples, the following regression model, described above, was fit to the observed scores:

$$y = \beta_0^* + \beta_u^* u + \beta_x^* x + \varepsilon^*$$

For each model, a two-tailed $t$-test of the form $t = \frac{\beta_x^*}{SE_{\beta_x^*}}$ was used to test the null hypothesis $H_0: \beta_x = 0$. The significance level of the $t$-test using $\alpha = 0.05$ was recorded for each of the 1,000 models in a given condition. The number of models for which the null hypothesis was rejected out of the 1,000 provided an empirical estimate of the Type I error rate for that condition.

In conditions for which the Type I error rate of the hypothesis test is inflated, the test is no longer valid and should not be used. A primary concern is thus with
determining not only how high Type I error rates are, but also whether the Type I error rate in a given condition is inflated or not. Even in conditions where the true Type I error rate of a test is correct (in this case 0.05), sampling error may cause the empirical (observed) Type I error rate to differ slightly from this value. Bradley’s (1978) “liberal” criterion was used to operationalize conditions resulting in inflated Type I error rates. According to this criterion, an observed Type I error rate above 0.075 constitutes a condition for which the true Type I error rate is considered to be inflated, hence invalidating the hypothesis test. In Parts I and II, a binary variable was recorded indicating whether the Type I error rates in each condition exceeded this criterion or not (1 = inflated, 0 = not inflated).\(^{14}\)

For Part III, additional parameter estimates for each model were recorded. These included the estimated standardized regression coefficient of \(x\), \(\beta_{zx}^*\), the observed change in \(R^2\) from adding \(x\) to the model, \(\Delta R_{x}^{*2}\), and the estimated standard error for \(\beta_{zx}^*\). These statistics are often used to characterize effect sizes, and provide a sense for how inferences regarding the practical significance (Kirk, 1996) as well as the statistical significance of \(x\) might be impacted when there is measurement error.

### 4.2.4 Data Analysis

Following recommendations by Zumbo and Harwell (1999), we analyzed the results of the simulation study as an experiment, using both statistical modeling and graphical analysis to better understand how Type I error rates vary as a function of the

\(^{14}\) Other options for determining whether a condition is evidence of truly inflated Type I error rates include the construction of confidence regions (e.g., Serlin, 2000) or using hypothesis tests of the null hypothesis \(H_0: \pi = 0.05\), where \(\pi\) is the true probability of a Type I error. Hypothesis testing yields a similar cutoff for significantly different from 0.05, and Serlin’s approach also requires specifying an (arbitrary) cutoff value to construct intervals. Hence the 0.075 criterion was used for continuity in the literature and for simplicity of interpretation.
independent variables. In Parts I and II, OLS regression (Cohen et al., 2003) was used to model observed Type I error rates as a function of the experimental factors, and a binary logistic regression (Agresti, 2007) was used to model the binary indicator of inflated/not inflated Type I error rates as a function of the experimental factors. These models were used as a tool to order the relative importance of the different experimental factors in affecting Type I error rates, and to facilitate interpretation of the results. For Part III, descriptive statistics of the outcomes were computed without the use of statistical modeling.

After estimating the linear regression model, the Pratt index (Thomas, Hughes, & Zumbo, 1998), was used to determine which factors were most important in affecting Type I error rates for Parts I and II. The Pratt index is a measure of relative variable importance that uniquely partitions the $R^2$ in a linear regression model among the independent variables. The Pratt index values are normed to sum to 1.0, and represent the percentage of a model $R^2$ uniquely attributable to each predictor. A generalization of the Pratt index, described by Thomas et al. (2008), was used to order the factors in the logistic regression models for Parts I and II. This generalization decomposes the pseudo-$R^2$ value described for logistic regression by McKelvey and Zavoina (1975) and Laitila (1993). Significance tests for these models were not used to guide interpretation, because the sample sizes (5,000 and 2,000) were large enough that most effects were statistically significant.

Both linear and logistic regression models were fit with main effects and all 2-way interactions; higher order interactions were not included to keep results manageable. This resulted in 21 estimated effects, plus an intercept term, in each model. To further
restrict attention to the most important factors, we focused on all variables meeting Thomas’ (1992) criteria that any variable with a normed Pratt index greater than $1/2p = 1/(2*21) = 0.024 = 2.4\%$, where $p$ is the number of predictors, be considered practically important in the model. This criterion was applied to importance values before rounding to three decimal points in the results section. If no interaction terms met this threshold, a model with only main effects was estimated instead. In order to avoid problems with multicollinearity of the interaction terms, all predictors were mean-centered prior to model estimation for both the linear and logistic models. For the logistic model, the predictors were also standardized.

Graphical representations were used to facilitate interpretation of the results, particularly interactions between factors. This is also in keeping with recommendations by Gelman et al. (2002) who advocated increasing the use of visual data displays to present tabular information. We follow data visualization principles described in various sources (e.g., Cleveland, 1993; Tufte, 1983; Wilkinson, 2005) and made possible with recent R packages (Sarkar, 2008; Wickham, 2009).

After identifying the most important experimental factors, the direction and magnitude of the standardized $\beta$-weights were used to interpret the effect of each factor. Modeling both the absolute Type I error rates and the occurrence of inflation allowed us to answer slightly different questions. The linear regression model addressed the question, “which factors are most important in affecting Type I error rates?” The logistic regression model addressed the slightly different question, “which factors are most important in determining whether Type I error rates will be inflated?” Hence we could use these results to answer the question of whether the reliability of $y$ seems to be
important in determining the Type I error rates of the test and also whether it is important relative to other factors in leading to a scenario where the hypothesis test is no longer valid.

4.3 Results

4.3.1 Part I Results

Table 4 shows summary statistics for the 5,000 randomly generated experimental conditions and resulting Type I error rates in Part I. Using the cutoff criteria of 0.075, 3,184 of the 5,000 conditions (63.68%) had inflated Type I error rates in Part I. Statistics for the design conditions were as expected. Although not apparent in the summary statistics, a histogram clearly shows that the distribution of observed Type I error rates was bi-modal; Figure 4 shows a histogram of the frequency of different observed Type I error rates. The vertical dashed line is placed at the cutoff of 0.075. In Part I, although many conditions led to Type I error rates below this value, there were a large number of Type I error rates at or near 1.0 (723, or 14.46% of cases greater than 0.975), suggesting that in many cases that are theoretically possible the test is almost guaranteed to give an incorrect answer and lead to a false positive finding.

Table 4. Summary Statistics for 5,000 Experimental Conditions in Part I.

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$rel(x)$</td>
<td>5000</td>
<td>0.499</td>
<td>0.291</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$rel(u)$</td>
<td>5000</td>
<td>0.498</td>
<td>0.287</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$rel(y)$</td>
<td>5000</td>
<td>0.501</td>
<td>0.291</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$Cor(\tau_x, \tau_u)$</td>
<td>5000</td>
<td>0.503</td>
<td>0.290</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$R^2$</td>
<td>5000</td>
<td>0.504</td>
<td>0.290</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Sample Size</td>
<td>5000</td>
<td>507.380</td>
<td>288.637</td>
<td>10</td>
<td>999</td>
</tr>
<tr>
<td>Type I Errors</td>
<td>5000</td>
<td>0.354</td>
<td>0.367</td>
<td>0.033</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Figure 4. Histogram of Observed Type I Error Rates in Part I.

The results of the linear model with all two-way interactions for Part I are presented in Table 5. Factors with a Pratt index greater than $1/2p = 1/42 = 0.024$ are indicated in boldface above a dashed line. The seven effects above the dashed line account for 89.3% of the model $R^2$, which is 0.7788, suggesting that they explain the vast majority of the predicted variance in observed Type I error rates. The effect of the correlation between the predictors ($Corxu$ in the table) is clearly the most important factor in the model, accounting for nearly 30% of the model $R^2$. The second most important factor is the reliability of $u$ ($Relu$), the predictor we are not testing, but which is related to $y$. Note, however, that the interaction between these two variables is also an important factor, and we cannot interpret the main effects without taking into account the interaction. The magnitude and direction of these effects suggest that an increase in the correlation between predictors will increase the Type I error rate, while an increase in the reliability of $u$ will decrease the Type I error rate. These are consistent with mathematical
predictions discussed earlier. However, we can also see that the effect of Corxu will decrease as RelU increases since the interaction is negative. Corxu, RelU, and their interaction account for 51.6% of the model $R^2$. Of the remaining important factors, the reliability of $x$ (RelX) appears to be the most important, while sample size (Sample.N) is the least important, although their importance is not substantially different. These effects are all positive, suggesting that all else being equal, an increase in the reliability of $x$ (the predictor we are testing), an increase in sample size, an increase in the reliability of $y$, or an increase in the true model $R^2$ will all increase the Type I error rate.

Table 5. Variable Ordering Based on Linear Model in Part I.

<table>
<thead>
<tr>
<th>Factor</th>
<th>$b$</th>
<th>$\beta$</th>
<th>SE</th>
<th>$t$</th>
<th>VIF</th>
<th>$r$</th>
<th>Pratt Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corxu</td>
<td>0.604</td>
<td>0.477</td>
<td>0.008</td>
<td>71.519</td>
<td>1.003</td>
<td>0.488</td>
<td>0.299</td>
</tr>
<tr>
<td>RelU</td>
<td>-0.489</td>
<td>-0.382</td>
<td>0.009</td>
<td>-57.225</td>
<td>1.005</td>
<td>-0.381</td>
<td>0.187</td>
</tr>
<tr>
<td>RelX</td>
<td>0.353</td>
<td>0.280</td>
<td>0.008</td>
<td>41.984</td>
<td>1.003</td>
<td>0.293</td>
<td>0.105</td>
</tr>
<tr>
<td>Rsq</td>
<td>0.351</td>
<td>0.277</td>
<td>0.008</td>
<td>41.483</td>
<td>1.003</td>
<td>0.273</td>
<td>0.097</td>
</tr>
<tr>
<td>RelY</td>
<td>0.352</td>
<td>0.280</td>
<td>0.008</td>
<td>41.851</td>
<td>1.004</td>
<td>0.258</td>
<td>0.093</td>
</tr>
<tr>
<td>Sample.N</td>
<td>0.000</td>
<td>0.240</td>
<td>0.000</td>
<td>35.863</td>
<td>1.005</td>
<td>0.267</td>
<td>0.082</td>
</tr>
<tr>
<td>RelU:Corxu</td>
<td>-0.671</td>
<td>-0.153</td>
<td>0.029</td>
<td>-22.847</td>
<td>1.005</td>
<td>-0.149</td>
<td>0.029</td>
</tr>
<tr>
<td>RelX:Corxu</td>
<td>0.505</td>
<td>0.116</td>
<td>0.029</td>
<td>17.347</td>
<td>1.004</td>
<td>0.134</td>
<td>0.020</td>
</tr>
<tr>
<td>Corxu:Rsq</td>
<td>0.493</td>
<td>0.112</td>
<td>0.029</td>
<td>16.807</td>
<td>1.007</td>
<td>0.112</td>
<td>0.016</td>
</tr>
<tr>
<td>RelU:Corxu</td>
<td>0.444</td>
<td>0.102</td>
<td>0.029</td>
<td>15.241</td>
<td>1.008</td>
<td>0.101</td>
<td>0.013</td>
</tr>
<tr>
<td>Corxu:Sample.N</td>
<td>0.000</td>
<td>0.087</td>
<td>0.000</td>
<td>13.080</td>
<td>1.005</td>
<td>0.103</td>
<td>0.012</td>
</tr>
<tr>
<td>RelX:RelU</td>
<td>-0.351</td>
<td>-0.081</td>
<td>0.029</td>
<td>-12.038</td>
<td>1.006</td>
<td>-0.085</td>
<td>0.009</td>
</tr>
<tr>
<td>RelU:RelY</td>
<td>-0.339</td>
<td>-0.078</td>
<td>0.029</td>
<td>-11.624</td>
<td>1.007</td>
<td>-0.070</td>
<td>0.007</td>
</tr>
<tr>
<td>RelY:Rsq</td>
<td>0.344</td>
<td>0.080</td>
<td>0.029</td>
<td>11.958</td>
<td>1.004</td>
<td>0.064</td>
<td>0.007</td>
</tr>
<tr>
<td>RelU:Sample.N</td>
<td>0.000</td>
<td>-0.074</td>
<td>0.000</td>
<td>-11.132</td>
<td>1.006</td>
<td>-0.063</td>
<td>0.006</td>
</tr>
<tr>
<td>RelU:Rsq</td>
<td>-0.270</td>
<td>-0.061</td>
<td>0.030</td>
<td>-9.119</td>
<td>1.006</td>
<td>-0.055</td>
<td>0.004</td>
</tr>
<tr>
<td>RelX:Rsq</td>
<td>0.190</td>
<td>0.043</td>
<td>0.029</td>
<td>6.509</td>
<td>1.005</td>
<td>0.062</td>
<td>0.004</td>
</tr>
<tr>
<td>Rsq:Sample.N</td>
<td>0.000</td>
<td>0.043</td>
<td>0.000</td>
<td>6.429</td>
<td>1.008</td>
<td>0.054</td>
<td>0.003</td>
</tr>
<tr>
<td>RelX:RelY</td>
<td>0.210</td>
<td>0.049</td>
<td>0.029</td>
<td>7.261</td>
<td>1.007</td>
<td>0.045</td>
<td>0.003</td>
</tr>
<tr>
<td>RelY:Sample.N</td>
<td>0.000</td>
<td>0.050</td>
<td>0.000</td>
<td>7.406</td>
<td>1.006</td>
<td>0.038</td>
<td>0.002</td>
</tr>
<tr>
<td>RelX:Sample.N</td>
<td>0.000</td>
<td>0.032</td>
<td>0.000</td>
<td>4.822</td>
<td>1.005</td>
<td>0.045</td>
<td>0.002</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.353</td>
<td>0.000</td>
<td>0.002</td>
<td>144.198</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Note: $R^2 = 0.7788$. Corxu = correlation between predictors; RelU = reliability of $u$; RelX = reliability of $x$; Rsq = true model $R^2$; RelY = reliability of $y$; Sample.N = sample size; “:” indicates an interaction term; VIF = Variance Inflation Factor; $b$ = unstandardized regression coefficient; $\beta$ = standardized regression coefficient. Pratt index values sum to 1.0 and hence represent unique proportions (percentages) of the total model $R^2$. The important factors, selected using the criteria of 0.024 described in the methods section, are listed in boldface.
Table 6 shows the results of the logistic regression model for Part I, again with factors classified as important listed in boldface above the dashed line. The rank ordering and directionality of the effects classified as important are identical to those in the linear model above. Using a predicted probability cutoff of $\hat{r} = 0.5$, the overall classification accuracy rate of the model for predicting which conditions will lead to inflated Type I error rates is 90.42%, with a sensitivity of 0.9139 and specificity of 0.8871. The pseudo-$R^2$ for the model is 0.9971. The generalized Pratt index values suggest the correlation between predictors is again the most important factor, followed by the reliability of $u$. The interaction between these factors ($RelU:Corxu$) was the only important interaction effect. Directionality of the interaction again suggests that the impact of the correlation lessens as the reliability of $u$ increases. Rank ordering of the other main effects is the same, although the sample size now appears somewhat less important than the other main effects. The main effects of $Corxu$ and $RelU$ and their interaction account for 57.3% of the model $R^2$. 

Table 6. Variable Ordering Based on Logistic Model in Part I.

<table>
<thead>
<tr>
<th>Var</th>
<th>( \beta )</th>
<th>( SE )</th>
<th>( z )</th>
<th>Pratt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corxu</td>
<td>3.696</td>
<td>0.133</td>
<td>27.895</td>
<td>0.302</td>
</tr>
<tr>
<td>RelU</td>
<td>-3.173</td>
<td>0.122</td>
<td>-25.955</td>
<td>0.218</td>
</tr>
<tr>
<td>RelX</td>
<td>2.087</td>
<td>0.095</td>
<td>21.857</td>
<td>0.099</td>
</tr>
<tr>
<td>Rsq</td>
<td>1.980</td>
<td>0.094</td>
<td>20.988</td>
<td>0.084</td>
</tr>
<tr>
<td>RelY</td>
<td>1.976</td>
<td>0.095</td>
<td>20.795</td>
<td>0.078</td>
</tr>
<tr>
<td>Sample.N</td>
<td>1.553</td>
<td>0.087</td>
<td>17.861</td>
<td>0.059</td>
</tr>
<tr>
<td>RelX:Corxu</td>
<td>-1.570</td>
<td>0.095</td>
<td>-16.567</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Note: pseudo-\( R^2 = 0.9971 \). Corxu = correlation between predictors; RelU = reliability of \( u \); RelX = reliability of \( x \); Rsq = true model \( R^2 \); RelY = reliability of \( y \); Sample.N = sample size; “:” indicates an interaction term. Pratt index values sum to 1.0 and hence represent unique proportions of the total pseudo-\( R^2 \). The important factors, selected as described in the methods section, are listed in boldface.

Figure 5 is a simple slopes plot used to illustrate the nature of the interaction between the correlation of the predictors and the reliability of \( u \). Each point represents the observed Type I error rate in one of the 5,000 experimental conditions in Part I. Moving from left to right, panel A shows all conditions where reliability of \( u \) was on the range (0, 0.25], panel B shows the range (0.25, 0.50], panel C shows the range (0.50, 0.75], and panel D shows the range (0.75, 1.0). The \( x \)-axis of each panel depicts the correlation between \( \tau_x \) and \( \tau_u \), while the \( y \)-axis depicts the observed Type I error rate. The regression
line in each panel shows the model-predicted Type I error rates using the linear model summarized in Table 5. The line represents the predicted Type I error rate when reliability of \( x \) and \( y \) equal 0.5, \( R^2 \) equals 0.5, and sample size is 507 (the mean of each value across all conditions), but reliability of \( u \) is set equal to the mean value within only that panel. It is clear that as the reliability of \( u \) increases (from panel A through panel D), the relationship between \( Cor(\tau_x, \tau_u) \) and Type I error rates becomes significantly weaker, holding all other factors constant. Looking at the individual points, it is evident that when the reliability of \( u \) is increased, Type I error rates are generally lower.

![Figure 5. Simple Slopes for Linear Model in Part I.](image)

Figure 5 is similar to Figure 5, but shows the observed presence or absence of Type I error rate inflation for each condition, using the 0.075 cutoff described earlier. Each point again represents a single experimental condition. The \( y \)-axis shows presence or absence of Type I error inflation (1 = inflated Type I error rates, 0 = not inflated, with jitter added to the points to avoid overplotting). Each panel (A-D) represents the same ranges of reliability of \( u \) as in Figure 5. The \( x \)-axis represents the correlation between predictors. The solid line in each panel shows the predicted probability of Type I error
inflation using the logistic regression model in Table 6. Again, the model was evaluated at the overall mean reliability of \( x \) and \( y \), \( R^2 \), and sample size, but at the local panel mean of the reliability of \( u \). The interaction between the correlation of the predictors and the reliability of \( u \) causes the line to decrease in slope and shift to the right. The dashed line in each panel indicates the correlation of predictors needed to have a predicted probability of 0.5 that Type I error rates will be inflated. In panel A, a correlation of about 0.1 is needed for a 50% chance of Type I error inflation, while in panel D, a correlation of just below 0.6 is needed. Looking at the individual points, when the reliability of \( u \) is very low (panel A), the majority of conditions lead to inflated Type I error rates. However, even when the reliability of \( u \) is above 0.75 (panel D) there are still a substantial number of cases leading to inflated Type I error rates.

![Figure 6. Simple Slopes for Logistic Model in Part I.](image)

**4.3.2 Part II Results**

Table 7 shows summary statistics for the 2,000 experimental conditions in Part II along with summary statistics of the observed Type I error rates. Just over half (1,073, or 53.65%) of the 2,000 conditions led to inflated Type I error rates using the 0.075
criterion, even in this restricted range of conditions. Figure 7 shows a histogram of Type I error rates for Part II. As in Part I, the distribution has a second peak of Type I error rates at or near 1.0, although it is somewhat smaller than the peak in Part I, with only 66 (or 3.3%) of the 2,000 conditions having Type I error rates above 0.975.

Table 7. Summary Statistics for 2,000 Experimental Conditions in Part II.

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>rel(x)</td>
<td>2000</td>
<td>0.798</td>
<td>0.116</td>
<td>0.601</td>
<td>1.000</td>
</tr>
<tr>
<td>rel(u)</td>
<td>2000</td>
<td>0.800</td>
<td>0.117</td>
<td>0.600</td>
<td>1.000</td>
</tr>
<tr>
<td>rel(y)</td>
<td>2000</td>
<td>0.799</td>
<td>0.115</td>
<td>0.600</td>
<td>1.000</td>
</tr>
<tr>
<td>Cor(τ_x, τ_u)</td>
<td>2000</td>
<td>0.373</td>
<td>0.214</td>
<td>0.000</td>
<td>0.750</td>
</tr>
<tr>
<td>R^2</td>
<td>2000</td>
<td>0.376</td>
<td>0.215</td>
<td>0.000</td>
<td>0.750</td>
</tr>
<tr>
<td>Sample Size</td>
<td>2000</td>
<td>508.000</td>
<td>284.661</td>
<td>11</td>
<td>998</td>
</tr>
<tr>
<td>Type I Errors</td>
<td>2000</td>
<td>0.221</td>
<td>0.268</td>
<td>0.029</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Figure 7. Histogram of Observed Type I Error Rates in Part II.
Table 8 shows the variable ordering of the factors based on the linear model for the data in Part II, with important factors listed in boldface above the dashed line. Nine factors are classified as important in this model, as opposed to seven in Part I. Again the correlation between $\tau_x$ and $\tau_u$ (Corxu) and the reliability of $u$ (RelU) are the most important factors, accounting for nearly half (49.7%) of the model $R^2$, which is 0.9028. Again, higher correlations lead to increased Type I error rates, while increasing reliability of $u$ leads to decreased Type I error rates; the interaction is also important, and again suggests that the effect of the correlation decreases as the reliability of $u$ increases. Although $R^2$ and sample size are still important main effects (with both again increasing Type I error rates), the reliability of $x$ and $y$ are not important factors in this model, as they were in Part I. In contrast to Part I, the interaction between Corxu and RelU is now the fourth most important factor, whereas it was the least important of those considered important in Part I. Also, there are four additional interaction effects that are classified as important compared to Part I. In fact, all possible interactions with Corxu and RelU and the other important main effects are classified as important. All interactions involving RelU are negative, which means that as RelU increases, the effect of the other factors becomes less important in affecting Type I error rates. Conversely, all interactions with Corxu are positive, suggesting that as the correlation between the predictors increases, the effect of the other factors (aside from RelU) have a stronger impact on increasing Type I error rates. These interactions account for 25.5% of the model $R^2$. 
Table 8. Variable Ordering Based on Linear Model in Part II.

<table>
<thead>
<tr>
<th>Var</th>
<th>$b$</th>
<th>$\beta$</th>
<th>SE</th>
<th>$t$</th>
<th>$p$</th>
<th>VIF</th>
<th>$r$</th>
<th>Pratt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coru</td>
<td>0.625</td>
<td>0.501</td>
<td>0.009</td>
<td>71.255</td>
<td>0.000</td>
<td>1.006</td>
<td>0.485</td>
<td>0.269</td>
</tr>
<tr>
<td>RelU</td>
<td>-1.015</td>
<td>-0.443</td>
<td>0.016</td>
<td>-62.894</td>
<td>0.000</td>
<td>1.009</td>
<td>-0.465</td>
<td>0.228</td>
</tr>
<tr>
<td>Rsq</td>
<td>0.428</td>
<td>0.344</td>
<td>0.009</td>
<td>48.778</td>
<td>0.000</td>
<td>1.010</td>
<td>0.330</td>
<td>0.126</td>
</tr>
<tr>
<td>RelU:Coru</td>
<td>-3.157</td>
<td>-0.297</td>
<td>0.075</td>
<td>-42.183</td>
<td>0.000</td>
<td>1.009</td>
<td>-0.305</td>
<td>0.100</td>
</tr>
<tr>
<td>Sample.N</td>
<td>0.000</td>
<td>0.266</td>
<td>0.000</td>
<td>37.835</td>
<td>0.000</td>
<td>1.007</td>
<td>0.281</td>
<td>0.083</td>
</tr>
<tr>
<td>Coru:Rsq</td>
<td>1.333</td>
<td>0.231</td>
<td>0.041</td>
<td>32.808</td>
<td>0.000</td>
<td>1.010</td>
<td>0.214</td>
<td>0.055</td>
</tr>
<tr>
<td>RelU:Rsq</td>
<td>-1.948</td>
<td>-0.182</td>
<td>0.075</td>
<td>-25.809</td>
<td>0.000</td>
<td>1.016</td>
<td>-0.190</td>
<td>0.038</td>
</tr>
<tr>
<td>Coru:Sample.N</td>
<td>0.001</td>
<td>0.168</td>
<td>0.000</td>
<td>23.888</td>
<td>0.000</td>
<td>1.012</td>
<td>0.174</td>
<td>0.033</td>
</tr>
<tr>
<td>RelU:Sample.N</td>
<td>-0.001</td>
<td>-0.146</td>
<td>0.000</td>
<td>-20.767</td>
<td>0.000</td>
<td>1.012</td>
<td>-0.179</td>
<td>0.029</td>
</tr>
<tr>
<td>Rsq:Sample.N</td>
<td>0.000</td>
<td>0.108</td>
<td>0.000</td>
<td>15.367</td>
<td>0.000</td>
<td>1.010</td>
<td>0.098</td>
<td>0.012</td>
</tr>
<tr>
<td>RelX</td>
<td>0.185</td>
<td>0.080</td>
<td>0.016</td>
<td>11.384</td>
<td>0.000</td>
<td>1.010</td>
<td>0.094</td>
<td>0.008</td>
</tr>
<tr>
<td>RelY</td>
<td>0.177</td>
<td>0.076</td>
<td>0.016</td>
<td>10.770</td>
<td>0.000</td>
<td>1.008</td>
<td>0.074</td>
<td>0.006</td>
</tr>
<tr>
<td>RelX:Coru</td>
<td>0.693</td>
<td>0.063</td>
<td>0.077</td>
<td>9.000</td>
<td>0.000</td>
<td>1.012</td>
<td>0.070</td>
<td>0.005</td>
</tr>
<tr>
<td>RelX:RelU</td>
<td>-0.975</td>
<td>-0.050</td>
<td>0.138</td>
<td>-7.059</td>
<td>0.000</td>
<td>1.010</td>
<td>-0.048</td>
<td>0.003</td>
</tr>
<tr>
<td>RelY:Coru</td>
<td>0.373</td>
<td>0.035</td>
<td>0.076</td>
<td>4.910</td>
<td>0.000</td>
<td>1.011</td>
<td>0.049</td>
<td>0.002</td>
</tr>
<tr>
<td>RelY:Sample.N</td>
<td>0.000</td>
<td>0.027</td>
<td>0.000</td>
<td>3.855</td>
<td>0.000</td>
<td>1.009</td>
<td>0.048</td>
<td>0.002</td>
</tr>
<tr>
<td>RelU:RelY</td>
<td>-0.651</td>
<td>-0.033</td>
<td>0.140</td>
<td>-4.659</td>
<td>0.000</td>
<td>1.008</td>
<td>-0.019</td>
<td>0.001</td>
</tr>
<tr>
<td>RelX:Rsq</td>
<td>0.308</td>
<td>0.029</td>
<td>0.076</td>
<td>4.063</td>
<td>0.000</td>
<td>1.011</td>
<td>0.018</td>
<td>0.001</td>
</tr>
<tr>
<td>RelX:Sample.N</td>
<td>0.000</td>
<td>0.021</td>
<td>0.000</td>
<td>3.013</td>
<td>0.000</td>
<td>1.007</td>
<td>0.026</td>
<td>0.001</td>
</tr>
<tr>
<td>RelX:RelY</td>
<td>0.201</td>
<td>0.010</td>
<td>0.140</td>
<td>1.431</td>
<td>0.153</td>
<td>1.011</td>
<td>0.040</td>
<td>0.000</td>
</tr>
<tr>
<td>RelY:Rsq</td>
<td>0.390</td>
<td>0.035</td>
<td>0.078</td>
<td>5.015</td>
<td>0.000</td>
<td>1.012</td>
<td>0.003</td>
<td>0.000</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.219</td>
<td>0.000</td>
<td>0.002</td>
<td>116.484</td>
<td>0.000</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Note: $R^2 = 0.9028$. All variables centered before fitting model. Coru = correlation between predictors; RelU = reliability of $u$; RelX = reliability of $x$; Rsq = true model $R^2$; RelY = reliability of $y$; Sample.N = sample size; ";:;" indicates an interaction term; VIF = Variance Inflation Factor. Pratt index values sum to 1.0 and hence represent unique proportions (percentages) of the total model $R^2$. The important factors, selected using the criteria of 0.024 described in the methods section, are listed in boldface.

Table 9 shows the results of the variable ordering for Part II based on the logistic regression model predicting inflation in Type I error rates. The pseudo-$R^2$ is 0.9986. The variable ordering results are very similar to those for the linear model for Part II, with the following differences. Sample size is slightly more important, and now ranked above all of the interaction terms considered important. There are now eight effects considered important, rather than nine in the linear model; the interaction between the correlation of
the predictors and sample size is no longer important. The other interactions and main effects have the same directionality and interpretation, but the rank ordering of the interactions is slightly different. The interactions account for slightly less than 17.3% of the pseudo-$R^2$. The correlation of the predictors and the reliability of $u$, together with their interaction, account for 66.5% of the model $R^2$.

Table 9. Variable Ordering Based on Logistic Model in Part II.

<table>
<thead>
<tr>
<th>Var</th>
<th>$\beta$</th>
<th>SE</th>
<th>$z$</th>
<th>$p$</th>
<th>Pratt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corxu</td>
<td>4.745</td>
<td>0.270</td>
<td>17.595</td>
<td>0.000</td>
<td>0.312</td>
</tr>
<tr>
<td>RelU</td>
<td>-4.313</td>
<td>0.248</td>
<td>-17.369</td>
<td>0.000</td>
<td>0.277</td>
</tr>
<tr>
<td>Rsq</td>
<td>2.971</td>
<td>0.182</td>
<td>16.350</td>
<td>0.000</td>
<td>0.123</td>
</tr>
<tr>
<td>Sample.N</td>
<td>2.433</td>
<td>0.165</td>
<td>14.750</td>
<td>0.000</td>
<td>0.089</td>
</tr>
<tr>
<td>RelU:Corxu</td>
<td>-2.265</td>
<td>0.180</td>
<td>-12.553</td>
<td>0.000</td>
<td>0.076</td>
</tr>
<tr>
<td>RelU:Rsq</td>
<td>-1.657</td>
<td>0.163</td>
<td>-10.183</td>
<td>0.000</td>
<td>0.041</td>
</tr>
<tr>
<td>Corxu:Rsq</td>
<td>1.588</td>
<td>0.167</td>
<td>9.517</td>
<td>0.000</td>
<td>0.032</td>
</tr>
<tr>
<td>RelU:Sample.N</td>
<td>-1.206</td>
<td>0.149</td>
<td>-8.099</td>
<td>0.000</td>
<td>0.024</td>
</tr>
<tr>
<td>Corxu:Sample.N</td>
<td>0.958</td>
<td>0.138</td>
<td>6.957</td>
<td>0.000</td>
<td>0.013</td>
</tr>
<tr>
<td>Rsq:Sample.N</td>
<td>0.678</td>
<td>0.121</td>
<td>5.586</td>
<td>0.000</td>
<td>0.006</td>
</tr>
<tr>
<td>RelX</td>
<td>0.476</td>
<td>0.111</td>
<td>4.292</td>
<td>0.000</td>
<td>0.004</td>
</tr>
<tr>
<td>RelY</td>
<td>0.440</td>
<td>0.109</td>
<td>4.035</td>
<td>0.000</td>
<td>0.003</td>
</tr>
<tr>
<td>RelX:RelU</td>
<td>-0.337</td>
<td>0.136</td>
<td>-2.471</td>
<td>0.013</td>
<td>0.001</td>
</tr>
<tr>
<td>RelX:Corxu</td>
<td>0.196</td>
<td>0.140</td>
<td>1.405</td>
<td>0.160</td>
<td>0.001</td>
</tr>
<tr>
<td>RelX:Rsq</td>
<td>-0.083</td>
<td>0.104</td>
<td>-0.794</td>
<td>0.427</td>
<td>0.000</td>
</tr>
<tr>
<td>RelY:Corxu</td>
<td>0.042</td>
<td>0.129</td>
<td>0.328</td>
<td>0.743</td>
<td>0.000</td>
</tr>
<tr>
<td>RelY:Sample.N</td>
<td>0.003</td>
<td>0.103</td>
<td>0.028</td>
<td>0.978</td>
<td>0.000</td>
</tr>
<tr>
<td>RelU:RelY</td>
<td>-0.060</td>
<td>0.131</td>
<td>-0.454</td>
<td>0.650</td>
<td>0.000</td>
</tr>
<tr>
<td>RelY:Sample.N</td>
<td>-0.111</td>
<td>0.105</td>
<td>-1.060</td>
<td>0.289</td>
<td>0.000</td>
</tr>
<tr>
<td>RelX:RelY</td>
<td>-0.042</td>
<td>0.095</td>
<td>-0.437</td>
<td>0.662</td>
<td>0.000</td>
</tr>
<tr>
<td>RelY:Rsq</td>
<td>0.099</td>
<td>0.108</td>
<td>0.917</td>
<td>0.359</td>
<td>0.000</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>2.956</td>
<td>0.215</td>
<td>13.742</td>
<td>0.000</td>
<td>--</td>
</tr>
</tbody>
</table>

Note: pseudo-$R^2 = 0.9986$. All variables standardized before fitting the logistic regression model. Corxu = correlation between predictors; RelU = reliability of $u$; RelX = reliability of $x$; Rsq = true model $R^2$; RelY = reliability of $y$; Sample.N = sample size; ‘:** indicates an interaction term. Pratt index values sum to 1.0 and hence represent unique proportions (percentages) of the total model $R^2$. The important factors, selected using the criteria of 0.024 described in the methods section, are listed in boldface.

Figure 8 shows the observed Type I error rates across all 2,000 experimental conditions in Part II. The figure visualizes the relationship between the three most
important factors \((\text{Cor}(\tau_x, \tau_u), \text{reliability of } u, \text{ and } R^2)\) and is organized to show their interactions. Moving from panel A through panel D, conditions with differing ranges of the reliability of \(u\) are shown. Within each panel, the y-axis represents observed Type I error rate, the x-axis represents \(\text{Cor}(\tau_x, \tau_u)\), and the shape of the points indicates the \(R^2\) (grouped into three ranges). The \(R^2\) for all x’s was between 0.5 and 0.75, between 0.25 and 0.75 for the triangles, and less than 0.25 for the circles. Within each panel a smoothed line was fit to the subgroups based on the \(R^2\) to give a sense for the overall trend in those conditions. For example, the x’s in panel A represent all conditions in which the reliability of \(u\) ranged from 0.6 to 0.7 and the \(R^2\) ranged from 0.5 to 0.75, and the smoothed line (dash-dot-dash) suggests a very strong positive relationship between the correlation of the predictors and subsequent Type I error rates.

Even for cases in panel A for which the \(R^2\) was less than 0.25, there is a substantial positive relationship between \(\text{Cor}(\tau_x, \tau_u)\) and Type I error rates, particularly for values of \(\text{Cor}(\tau_x, \tau_u)\) greater than about 0.25. In contrast, when the reliability of \(u\) is greater than 0.9 (panel D), there is very little inflation of Type I error rates, with only a few exceptions, mostly when the value of \(\text{Cor}(\tau_x, \tau_u)\) is above 0.6. Here, all three loess lines have a nearly flat slope, indicating little relationship between \(\text{Cor}(\tau_x, \tau_u)\) and Type I error rates, aside from the few cases noted. Within each panel, the differing loess lines demonstrate the interaction between \(\text{Cor}(\tau_x, \tau_u)\) and \(R^2\) - as \(R^2\) increases, so too does the relationship between \(\text{Cor}(\tau_x, \tau_u)\) and Type I error rates. Similarly, moving across panels, from A to D, shows the interaction of these two factors with the reliability of \(u\) – as the reliability of \(u\) increases, the effect of both \(\text{Cor}(\tau_x, \tau_u)\) and \(R^2\) diminishes.
Figure 8. Observed Type I Error Rates by Reliability of $u$ and R-square in Part II.

Note: Each point represents 1 of the 2,000 experimental conditions in Part II. Each panel displays conditions with the indicated ranges of reliability of $u$. Plotting symbols denote the true model R-square for each condition. The lines within each plot are smoothed loess lines, with each line type representing a different range of R-squared.
4.3.3 Part III Results

Table 10 shows the observed Type I error rate across all 96 conditions studied in Part III. Overall, 38 (39.6%) of the conditions had inflated Type I error rates using the 0.075 criteria; these conditions are listed in boldface. The values in parentheses below row and column headers denote the expected value of the correlations between observed scores $x$ and $u$, and the $R^2$ when the observed scores of $y$ are regressed on observed scores of $u$, which will be attenuated due to the measurement error. These values give a sense for what each condition might look like in practice. There are two important points to take away from this table. First, when the predictors are uncorrelated (the first 4 rows), there is no inflation in Type I error rates. This supports earlier simulation results (Brunner & Austin, 2009) and mathematical derivations (Carroll et al., 2006; Gustafson, 2004). The second is based on an examination of the last 4 rows of the table. The Type I error rates in this section can be highly inflated, suggesting that even in a case where the observed correlation between two predictors is 0.56, and the model $R^2$ is no higher than 0.32, Type I error rates could be as high as 0.954! Moreover, Type I error rates are inflated (hence invalidating the hypothesis test) in 13 out of the 16 conditions in the last four rows.
Table 10. Observed Type I Error Rates in Part III.

<table>
<thead>
<tr>
<th>$\text{Cor}(\tau_x, \tau_u)$</th>
<th>Sample Size</th>
<th>$R$-square</th>
<th>$0.01961$</th>
<th>$0.1304$</th>
<th>$0.2539$</th>
<th>$0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.013)</td>
<td>(0.083)</td>
<td>(0.162)</td>
<td>(0.320)</td>
</tr>
<tr>
<td>$0.0$</td>
<td>50</td>
<td>0.055</td>
<td>0.05</td>
<td>0.052</td>
<td>0.057</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.05</td>
<td>0.043</td>
<td>0.031</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.061</td>
<td>0.046</td>
<td>0.04</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>$0.1$</td>
<td>50</td>
<td>0.055</td>
<td>0.059</td>
<td>0.043</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.08)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.043</td>
<td>0.058</td>
<td>0.056</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.041</td>
<td>0.057</td>
<td>0.041</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.049</td>
<td>0.045</td>
<td>0.051</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td>$0.3$</td>
<td>50</td>
<td>0.048</td>
<td>0.044</td>
<td>0.044</td>
<td>0.053</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.24)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.052</td>
<td>0.049</td>
<td>0.077</td>
<td>0.067</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.047</td>
<td>0.07</td>
<td>0.099</td>
<td>0.171</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.047</td>
<td>0.092</td>
<td>0.132</td>
<td>0.263</td>
<td></td>
</tr>
<tr>
<td>$0.5$</td>
<td>50</td>
<td>0.066</td>
<td>0.066</td>
<td>0.063</td>
<td>0.078</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.061</td>
<td>0.058</td>
<td>0.073</td>
<td>0.103</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.072</td>
<td>0.114</td>
<td>0.184</td>
<td>0.405</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.061</td>
<td>0.171</td>
<td>0.333</td>
<td>0.658</td>
<td></td>
</tr>
<tr>
<td>$0.6$</td>
<td>50</td>
<td>0.055</td>
<td>0.068</td>
<td>0.081</td>
<td>0.093</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.053</td>
<td>0.067</td>
<td>0.088</td>
<td>0.138</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.057</td>
<td>0.149</td>
<td>0.266</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.078</td>
<td>0.245</td>
<td>0.503</td>
<td>0.863</td>
<td></td>
</tr>
<tr>
<td>$0.7$</td>
<td>50</td>
<td>0.049</td>
<td>0.074</td>
<td>0.094</td>
<td>0.141</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.56)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.056</td>
<td>0.076</td>
<td>0.113</td>
<td>0.223</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.077</td>
<td>0.218</td>
<td>0.376</td>
<td>0.766</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.084</td>
<td>0.376</td>
<td>0.652</td>
<td>0.954</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Cells with inflated Type I error rates (>0.075) are listed in boldface. Values in parentheses are the expected value of the observed score (attenuated) correlations between $x$ and $u$, and squared multiple correlations between $y$ and $u$.

Table 11 shows the mean estimated standardized regression coefficient, $\beta^*_xz$, across all 96 conditions, and Table 12 shows the mean change in $R$-square resulting from adding $x$ into the regression model after $u$, $\Delta R^2_x$. Both are commonly used to determine a predictor’s importance in a regression model, with the change in $R$-square being a commonly recommended effect size. Because the true standardized regression
coefficient, $\beta_{zx}$, and the true change in $R^2$, $\Delta R^2_x$, are both 0, the values in Table 11 and Table 12 represent the bias in both of these estimates. For the standardized regression coefficient, the bias is worse in the same conditions for which the inflation in Type I errors is worse. In the top four rows (no correlation between predictors), the standardized coefficient is unbiased, whereas in the final four rows the estimated standardized regression coefficient can be as high as 0.117, certainly a non-trivial value. Estimates of $R^2$ are known to be positively biased even with no measurement error (Cohen et al., 2003), and the estimates in Table 12 are also positively biased in all cases. However, although the $\Delta R^2_x$ values are non-zero, they are not particularly large; they would never be characterized as more than “small” effects based on Cohen’s (1992) guidelines. While increasing correlation between predictors and $R^2$ appear to slightly increase the bias in the $\Delta R^2_x$ values, the effect does not appear strong. In addition, unlike Type I error rates and the standardized coefficient estimates, increasing sample size appears to reduce the bias in the $\Delta R^2_x$ values.

Interestingly, the bias in the standardized coefficients appears to be largely independent of sample size, while Type I error rates become worse as the sample size increases and the bias in the $\Delta R^2_x$ values decreases as sample size increases. For example, looking in the last four rows of the furthest right column of each table (correlation = 0.7, $R^2 = 0.5$), the bias in the standardized regression coefficient is steady at between 0.11 and 0.12. Meanwhile, Type I error rates increase from 0.141 to 0.954 as sample size increases from 50 to 1,000, but the $\Delta R^2_x$ decreases from 0.024 to 0.010, approaching the true score $\Delta R^2_x$ of 0. In at least some conditions then, increasing sample size will: severely inflate Type I error rates, leave bias in estimated regression
coefficients unchanged, and lead to more accurate estimates of one commonly used effect size ($\Delta R^2$). This clearly highlights that the effects of measurement error can be complex and can potentially affect different statistics of interest in opposite ways even in the same conditions.

Table 11. Mean Estimated Standardized Coefficient in Part III.

<table>
<thead>
<tr>
<th>$Cor(r_{xy}, r_{u})$</th>
<th>Sample Size</th>
<th>R-square 0.01961 (0.013)</th>
<th>R-square 0.1304 (0.083)</th>
<th>R-square 0.2539 (0.162)</th>
<th>R-square 0.5 (0.320)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0 (0.0)</td>
<td>50</td>
<td>-0.001</td>
<td>0.002</td>
<td>-0.005</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-0.002</td>
<td>0.001</td>
<td>0.002</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>0.1 (0.08)</td>
<td>50</td>
<td>0.004</td>
<td>0.008</td>
<td>0.005</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.001</td>
<td>0.012</td>
<td>0.004</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.002</td>
<td>0.005</td>
<td>0.006</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.001</td>
<td>0.005</td>
<td>0.005</td>
<td>0.010</td>
</tr>
<tr>
<td>0.3 (0.24)</td>
<td>50</td>
<td>0.000</td>
<td>0.020</td>
<td>0.021</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.006</td>
<td>0.021</td>
<td>0.022</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.008</td>
<td>0.017</td>
<td>0.028</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.007</td>
<td>0.019</td>
<td>0.026</td>
<td>0.035</td>
</tr>
<tr>
<td>0.5 (0.4)</td>
<td>50</td>
<td>0.021</td>
<td>0.036</td>
<td>0.049</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.017</td>
<td>0.032</td>
<td>0.043</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.015</td>
<td>0.035</td>
<td>0.046</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.012</td>
<td>0.034</td>
<td>0.048</td>
<td>0.067</td>
</tr>
<tr>
<td>0.6 (0.48)</td>
<td>50</td>
<td>0.023</td>
<td>0.048</td>
<td>0.071</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.019</td>
<td>0.045</td>
<td>0.068</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.018</td>
<td>0.045</td>
<td>0.064</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.017</td>
<td>0.045</td>
<td>0.064</td>
<td>0.090</td>
</tr>
<tr>
<td>0.7 (0.56)</td>
<td>50</td>
<td>0.025</td>
<td>0.050</td>
<td>0.090</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.011</td>
<td>0.058</td>
<td>0.081</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.023</td>
<td>0.059</td>
<td>0.080</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.022</td>
<td>0.060</td>
<td>0.083</td>
<td>0.116</td>
</tr>
</tbody>
</table>

*Note:* Values in parentheses are the expected value of the observed score (attenuated) correlations between $x$ and $u$, and squared multiple correlations between $y$ and $u$. 
Table 12. Mean Estimated Change in R-square in Part III.

<table>
<thead>
<tr>
<th>$Cor(r_x, r_u)$</th>
<th>Sample Size</th>
<th>0.01961 (0.013)</th>
<th>0.1304 (0.083)</th>
<th>0.2539 (0.162)</th>
<th>0.5 (0.320)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>50</td>
<td>0.020</td>
<td>0.019</td>
<td>0.018</td>
<td>0.015</td>
</tr>
<tr>
<td>(0.0)</td>
<td>100</td>
<td>0.010</td>
<td>0.009</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>0.1</td>
<td>50</td>
<td>0.020</td>
<td>0.019</td>
<td>0.017</td>
<td>0.013</td>
</tr>
<tr>
<td>(0.08)</td>
<td>100</td>
<td>0.010</td>
<td>0.010</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>0.3</td>
<td>50</td>
<td>0.020</td>
<td>0.020</td>
<td>0.018</td>
<td>0.015</td>
</tr>
<tr>
<td>(0.24)</td>
<td>100</td>
<td>0.010</td>
<td>0.010</td>
<td>0.010</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>0.5</td>
<td>50</td>
<td>0.022</td>
<td>0.021</td>
<td>0.019</td>
<td>0.017</td>
</tr>
<tr>
<td>(0.4)</td>
<td>100</td>
<td>0.011</td>
<td>0.010</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.001</td>
<td>0.002</td>
<td>0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>0.6</td>
<td>50</td>
<td>0.021</td>
<td>0.021</td>
<td>0.021</td>
<td>0.019</td>
</tr>
<tr>
<td>(0.48)</td>
<td>100</td>
<td>0.011</td>
<td>0.010</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.002</td>
<td>0.003</td>
<td>0.005</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.001</td>
<td>0.002</td>
<td>0.004</td>
<td>0.007</td>
</tr>
<tr>
<td>0.7</td>
<td>50</td>
<td>0.020</td>
<td>0.022</td>
<td>0.024</td>
<td>0.024</td>
</tr>
<tr>
<td>(0.56)</td>
<td>100</td>
<td>0.011</td>
<td>0.012</td>
<td>0.013</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.002</td>
<td>0.004</td>
<td>0.006</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.001</td>
<td>0.003</td>
<td>0.005</td>
<td>0.010</td>
</tr>
</tbody>
</table>

*Note:* Values in parentheses are the expected value of the observed score (attenuated) correlations between $x$ and $u$, and squared multiple correlations between $y$ and $u$. 
5 Conclusion

This chapter briefly addresses the four research questions presented at the end of Chapter 2. Implications of the findings for researchers and educators are discussed and potential solutions are referenced. Novel contributions of the studies in Chapters 3 and 4 are summarized. Finally, limitations and directions for future research are presented.

5.1 Study 1 Synopsis

In simple regression, measurement error can distort the standard errors and the variability of the estimated regression coefficient, as well as the estimated residual variance, but for the hypothesis test that \( \beta_x^* = 0 \), all of these impacts balance out. In the end, Type I error rates are left unchanged and the hypothesis test is a valid test under the null model, as has been pointed out in the statistical literature (Cochran, 1968; Carroll et al., 2006; Bunacoursi, 2010). In fact, this test is exactly correct assuming predictor and measurement errors are normally distributed and uncorrelated (Buonaccorsi, 2010), as they were in this simulation.

In addition to the finding that Type I error rates are correct, there are three messages to highlight. First, the standard errors and variance of the estimated parameters will, in general, be misleading relative to the variance in the parameters in which we are truly interested. Second, the estimated residual variance in the model will, in general, not be a good representation of the true residual variance if there is measurement error in \( y \). Finally, the impacts of measurement error can differ substantially depending upon whether one is interested in standardized or unstandardized regression coefficients. This last fact is not often discussed in the statistical literature. In educational and
psychological research, where standardized coefficients are often reported, it is important
to be aware of this difference.

There is also an important limitation to point out. Although the hypothesis test is
correct when $\beta_x = 0$, this will not necessarily be true when $\beta_x \neq 0$. Carroll et al. (2006)
show that when $\beta_x \neq 0$, the simple regression model has reduced power and yields a
biased estimate. In other words, these results do not suggest that measurement error will
never impact simple regression, only that Type I error rates will not be inflated when the
null hypothesis is true.

5.2 Study 2 Synopsis

To summarize Study 2 succinctly: random measurement error in the independent
variables can inflate Type I error rates, even in routinely encountered situations, and in
some cases by a very substantial amount. More specific questions are addressed below.

*How does measurement error in y impact the Type I error rates in multiple regression?*

Across the entire space of theoretically possible scenarios, all else being equal,
increasing the reliability of $y$ increases the Type I error rate. Put differently, when
comparing two studies, the one with a more reliable criterion ($y$) variable will have a
higher chance of committing a Type I error. One reason may be that measurement error
in $y$ increases the estimated residual variance, which in turn increases the standard errors
for the estimated $\beta^*$ coefficients, and reduces the $t$-statistics. In some sense, measurement
error in $y$ provides a buffer to the problem. In the restricted space studied in Part II, the
reliability of $y$ was not one of the important factors, suggesting that this is not likely to be
the most important factor determining the probability of a Type I error in practice.
Moreover, measurement error in $y$ can increase the bias of the estimated standardized
coefficients. These results should not suggest that it is better to have measurement error in \( y \); they simply highlight that Type I error rates may be more problematic in cases where the criterion variable is measured with very little error.

*Of the relevant factors, which are most important in causing Type I error rates to be inflated in multiple regression? Which factors lead to the greatest inflation of Type I error rates?*

Although the Type I error rate will depend on all six factors studied, it is clear that the correlation between the predictors is the most important factor, followed by the reliability of \( u \), the predictor we are *not* testing. Increasing the correlation increases Type I error rates while increasing the reliability of \( u \) brings Type I error rates back towards the nominal level. There was also an interaction between these two factors, such that increasing the reliability of \( u \) mitigated the impact of the correlation. This was true across the entire range of possible scenarios (Part I) and a more restricted space (Part II). Note that if there is no correlation between the predictors, or if the reliability of \( u \) is 1.0, there will be no inflation in Type I error rates. Hence these are both necessary, although not sufficient, conditions for inflated Type I error rates. This necessity may be one reason they consistently emerged as the most important factors in the results of Study 2. The results also highlighted that the relationship between the correlation and the reliability of \( u \) is complex, and that in the end Type I error rates will depend upon all six factors.

In Parts I and II of Study 2, increasing the remaining four factors (reliability of \( x \) and \( y \), \( R^2 \), and sample size) always led to an increase in Type I error rates, but their importance varied. Across the entire possible space (Part I), these four main effects were the only other factors considered important. Yet in Part II, focusing on the range of scenarios that might be considered more typical of applied research, the reliability of \( x \)
and $y$ did were not as important. One possible explanation for this difference is that when
the reliability of $x$ or $y$ can vary across the entire possible range (0 to 1), they may have a
strong impact on Type I error rates; however, once they reach a certain minimum
threshold (0.60 in this study) they are no longer likely to be the determining factor for
subsequent Type I error rates.

In the restricted range of Part II, $R^2$ and sample size continued to be important,
and Type I error rates also appear to be driven by the interactions between the correlation
of the predictors and the reliability of $u$ with these two factors. In a sense, the correlation
of the predictors and the reliability of $u$ can be seen as moderators of the effects of
sample size and $R^2$, with the correlation enhancing the effects and the reliability of $u$
diminishing the effects. Thus, in commonly encountered situations, we can conclude that
the correlation between predictors will drive the inflation in Type I error rates, both alone
and by enhancing the impact of sample size and model $R^2$, while increasing the reliability
of $u$ will drive Type I error rates back down, partially by diminishing the effect of the
other factors, and partially on its own.

Not surprisingly, the same factors that drive Type I error rates upwards are the
same factors that increase the probability of inflated Type I error rates, as evidenced by
the congruence between findings based on the linear regression and logistic regression
analyses performed in Parts I and II.

*What level of Type I error inflation could be anticipated in a realistic data analysis
scenario? How are measures of effect size impacted in this case?*

Looking at some specific scenarios that are commonly encountered in practice
(Part III) suggests that Type I error rates can be highly inflated, but the impact on effect
sizes is difficult to summarize or predict ahead of time. In over 1/3 of the conditions
studied, Type I error rates were inflated, sometimes as high as 0.954. Estimated standardized regression coefficients were noticeably biased, while the bias in the $\Delta R^2$ did not seem substantially worse than the usual bias in this estimator with no measurement error. The complexity of the overall problem and the difficulty inherent in predicting when it will be a problem was emphasized by the fact that across the same four conditions, we could reach the following potentially contradictory conclusions: Type I error rates will be drastically inflated, bias in the standardized regression coefficient will be unaffected (but can be quite high), and bias in the $\Delta R^2$ will diminish. The message to researchers is: be very wary of hypothesis tests in multiple regression when there is any measurement error in correlated predictors, and be careful in selecting other statistics, such as effect sizes, which may be impacted in complex ways.

5.3 Implications for Researchers

This problem has serious implications for applied researchers. Four implications are discussed here: existence of the problem, uninterpretability of common hypothesis tests, scope and magnitude of the problem, and the potential for false positives in the research literature. The first implication can be stated simply: measurement error can cause relationships or effects to appear larger than they really are and, in some cases, it can create observed effects where there are no true effects. In other words, contrary to the commonly held belief that measurement error will tend to cause a downward bias, in multiple regression, it can actually have the opposite effect. Worse still, many research design characteristics that we seek to maximize – such as $R^2$, reliability of focal predictors, reliability of the criterion, and sample size – may actually make the problem worse.
When conducted properly, and under the proper circumstances, the hypothesis test in multiple regression studied here answers the question, “if $H_0$ were true, what is the probability that I would observe a $\beta_x$ value this large or even larger?” This may or may not be useful (Cohen, 1994), but it does provide interpretable information. When the $p$-value is incorrect (too small), as it was for many cases studied in Chapter 4, the resulting answer becomes literally meaningless. In essence, when a researcher is interested in asking about $\beta_x$, but the independent variables are measured with error, the hypothesis test actually answers a question about $\beta^*_x$. This is akin to asking someone for the time, and having them respond with the time they were born – the units appear correct, and the answer may or may not be reasonable, but it simply is not what we wanted to know. The implication is that, in many cases, hypothesis tests based on multiple regression analyses are meaningless.

Hypothesis testing is designed to control the probability of rejecting true null hypotheses (Type I errors), which is done by setting the $\alpha$ level; when $p$-values are systematically too small, as they were here, researchers are no longer controlling the Type I error rate, and the purpose of hypothesis testing is nullified. The results in Part II suggest that following recommendations to direct attention towards effect sizes rather than $p$-values (Fan, 2001; Kirk, 1996; Wilkinson & Task Force on Statistical Inference, 1999) may be beneficial, it is difficult to make general statements. Moreover, effect sizes are often used after a decision about significance has been reached, which will be based on the incorrect hypothesis test.

While incorrect hypothesis tests are clearly problematic, matters are made worse by the magnitude of the inaccuracies found, and the potential scope of the problem. Multiple
regression is one of the most widely used data analysis techniques, but the problem may be even more widespread:

We have illustrated inflation of the Type I error rate for the normal linear model with simple additive measurement error, but the problem is much more general. We suggest that regardless of the type of measurement error and regardless of the statistical method used, ignoring measurement error in the independent variables can seriously inflate the Type I error rate" (Brunner & Austin, 2009, p. 40, italics in original).

While some qualifications to this statement need to be made (for example: error in $x$ only will not cause a problem; there is no problem if $x$ and $u$ are uncorrelated), Brunner and Austin go on to provide examples leading to inflated Type I error rates in other instantiations of the general linear model such as ANOVA, and Maxwell and Delaney (1993) also make this connection. Discussions of problems for ANCOVA cited earlier are another example. Although these studies and the present study were limited to main effects only (i.e., no interactions), there have also been demonstrations that measurement error in the independent variables can cause spurious interaction effects, framed from both regression and ANOVA contexts (Culpepper, 2012; Embretson, 1996; Kang & Waller, 2005). Recent interest has also focused on the impact that measurement error may have when statistical models are used to aid in causal analysis, for example through the use of propensity scores (Steiner, Cook, & Shadish, 2011).

Gustafson (2004) describes a “folk theorem” (p. 25) stating that problems in general linear models will often extend to generalized linear models, such as logistic regression. Although this is not strictly true, multiple authors have recently documented the potential for inflated Type I error rates in logistic regression when the independent variables are measured with error, mostly in the context of epidemiological and medical research (Austin & Brunner, 2004; Brunner & Austin, 2009; Fewell, Smith, & Sterne, 2007; Fraser & Stram, 2001).
In psychometrics, logistic regression is a commonly used method for detecting differential item functioning (DIF; Swaminathan & Rogers, 1990; Zumbo, 1999). In DIF studies, one tests the effect of a grouping variable \( (x, \text{perhaps perfectly reliable}) \) while including a total test score in the model \( (u, \text{includes measurement error}) \). If the results found here do generalize, then we would expect to see inflated Type I error rates for the test of the null hypothesis \( H_0: \beta_x = 0 \). Indeed, this is a widely recognized problem when using logistic regression to test for DIF (DeMars, 2010; Jodoin & Gierl, 2001; Li & Stout, 1996). Moreover, DeMars notes that, “the problem [of inflated Type I error rates] worsens as the test scores become less reliable or the sample size increases” (p. 962), which again exactly mirrors the results found here. Of course, the results here also suggest that this will only be a problem if \( \tau_x \) and \( \tau_u \) are correlated, which in the DIF context means the two groups differ in overall ability, and is referred to as impact. Again, this is exactly what has been found in the DIF literature. All of this suggests that a) researchers should be concerned about this problem in a variety of contexts, and b) developing an intuition for the problem is important so that researchers are aware of when it is relevant and can find ways to prevent or mitigate the problem.

These implications suggest researchers should use caution when interpreting hypothesis test results if independent variables are measured with error, and that changes to common practice may be warranted. Another problem involves prior research that has already been conducted, and the potential for false research findings in the published literature. Although inflated Type I error rates are problematic and invalidate the use of a hypothesis test, they do not immediately prove that research findings are false. As Cohen (1994) reminds us, whether a given relationship is actually true (i.e., whether \( H_0 \) is
actually false) is a separate question that cannot be answered by a hypothesis test. So inflating Type I error rates from 0.05 to 0.5 or even 0.95 does not mean that 50% or 95% of published research findings based on multiple regression are false.

Ioannidis (2005) provides an analysis of the question, “what is the probability that a given research finding is false?” Ioannidis suggests that a published research finding is more likely true than false if $(1 - \beta)R > \alpha$, where $\beta$ represents power, $R$ is the “ratio of the number of ‘true relationships’ to ‘no relationships’ among those tested in the field” (p. 0696), and $\alpha$ is the Type I error rate.$^{15}$ Clearly, as $\alpha$ increases, the chance that a published research finding represents a true finding diminishes.$^{16}$ Let us assume $R = 1$, meaning the chance that a tested relationship really exists is 50/50. This implies that a finding is more likely true than false if $(1 - \beta) > \alpha$. This inequality shows that, all else being equal, as the Type I error rate increases, the odds that a published research finding is true decrease. Hence, studies based on multiple regression models with independent variables measured with error will likely have a lower chance of being true than might be anticipated.

The question many researchers will be most interested in is: exactly when do I need to worry about this problem and what can be done? Unfortunately, an answer to the first part of the question is nearly impossible to provide. As shown in Study 2 and discussed in prior literature, it is immensely difficult to specify ahead of time the definite impacts of measurement error, and doing so often requires knowledge to which we may not have access (e.g., the exact reliability of variables or relationships among true scores).

---

$^{15}$ This is the simpler form of the relationship, and assumes there is no systematic bias in the reporting of research findings, which could make the problem even worse.

$^{16}$ Ioannidis cites the familiar intuition that measurement error will reduce observed relationships, and hence reduce the problem of false positives.
Compounding these complexities are problems of sampling variability – even if we had knowledge of the situation in the population, these may or may not provide an accurate sense of any given sample, particularly when considering properties of reliabilities and measurement errors (Zimmerman, 2011). The best course of action is probably to select methods that are less susceptible to measurement error and, perhaps more importantly, to develop an awareness and intuition about the problem for current and future researchers. These are discussed in the next two sections.

5.4 Potential Solutions

Numerous statistical methods exist for analyzing data that contain measurement error. Brunner and Austin (2009) advocate the use of latent variable models such as the structural equation models described by Jöreskog (1978) or Bollen (1989) and extended by others, such as Muthén (2002). Other options include errors in variables (EIV) models (e.g., Fuller, 1987), which have recently been recommended and tested in some limited scenarios similar to those studied in this thesis (Culpepper, 2012; Culpepper & Aguinis, 2011). Carroll et al. (2006) and Buonaccorsi (2010) provide extensive overviews of other options, and Gustafson (2004) provides a solution based on Bayesian methods for the situation considered here. Cohen et al. (2003, pp. 144–145) mention an ad-hoc solution that uses a matrix of disattenuated correlations to estimate the regression model, but this method has not been tested, and they note that standard errors and hypothesis tests may no longer be correct. Among other things, Zinbarg et al. (2010) recommend conducting sensitivity analyses to evaluate the potential for Type I errors in specific cases (similar recommendations have been made for partial correlations, e.g., Sechrest, 1963; Strauss,
1981), and also to use caution in the language used to present the results of analyses that are based on these situations.

While many of these methods provide elegant solutions to the problem, they face at least three obstacles. First, the methods often require more information or data than researchers have – for example, multiple measurements for each variable, or accurate knowledge about the reliability of the scores. Second, even in cases where the data are available, most of the methods require substantially greater levels of expertise than do standard multiple regression techniques. Third, researchers are not likely to even consider these methods if they are unaware of the problems. Measurement textbooks cover measurement error at great length, but spend little space discussing the biasing impacts on statistical analyses while statistics courses, which spend a great deal of time on bias and hypothesis testing, include little discussion of measurement error. These latter two obstacles highlight the importance of education in addressing the problem.

5.5 Implications for Teaching

When considering the implications for teaching statistics, or preparing researchers and methodologists in education and psychology, it is useful to apply the constructs of statistical literacy, statistical reasoning, and statistical thinking, as described for example by Garfield and Ben-Zvi (2007). Briefly, statistical literacy refers to the basic concepts and knowledge one needs to be an informed consumer of statistical information, statistical reasoning refers to the way people reason and think about statistical ideas (which could include heuristics and biases), and statistical thinking refers to the higher-order skills developed and applied by professional statisticians.
Regarding statistical literacy, it is probably most important that people know what measurement error is conceptually, when it is likely to arise, that it can create severely misleading statistical results, and that there are methods that do (and do not) address the problem. Cobb and Moore (1997) point out that statistics is fundamentally about applied problems in context – the context is what gives meaning to analyses and interpretations. Nearly all contexts will involve some sort of measurement error. Yet while curriculum guides, such as the AP Statistics Course Description (The College Board, 2010), the recently developed U.S. Common Core State Standards for Mathematics (National Governors Association Center for Best Practices, 2010), or the American Statistical Association’s GAISE College Report (Aliaga et al., 2010) all frame statistics as fundamentally concerned with variability, none of these documents contained the term “measurement error.” In line with Garfield et al. (2002), who do list “measurement error” as a topic in the “highest priority” section of their proposed course outline, and in light of the disastrous consequences ignoring measurement error can have, it seems essential to include this topic at all levels of the curriculum, even if only to raise awareness and define the concept.

Moving somewhat up the hierarchy to statistical reasoning, two concepts stand out from this study. First is the notion that measurement error can cause spurious effects and upward biases as well as downward biases that drown out a signal. This latter idea seems prevalent, both as an intuition, and in discussions of the impacts of measurement error (e.g., Baugh, 2002). Second is differentiating between “errors” that arise from variability in the phenomenon under study versus “errors” that arise from errors of measurement. More technically, this is distinguishing between the random or stochastic aspects that are
included in a model (for example the $\varepsilon$ term in the regression model) and the random portions that are not included, such as the measurement errors in $x$ or $u$ in the examples above. Measurement error models, such as those mentioned earlier in this section, account for measurement error by explicitly including the measurement errors in the model.

These are more subtle and advanced concepts than may be necessary to include in the curriculum at all levels, but certainly seem important for methodologists and researchers more generally. Hence, it would be good to include these concepts more substantially in graduate student methods training. Out of 628 pages in their text on regression, Cohen et al. (2003) devote fewer than 10 pages to the effects of measurement error, and do not discuss the potential for Type I error inflation. Moreover, the examples in the book analyze data that contain measurement error without taking it into account. Henson et al. (2010) write that, “Doctoral preparation is a complex enterprise, to be sure, but mastery of quantitative methodologies is a key part of that complexity that affects not only data analysis but the very conceptualization of research questions” (p. 230). Clearly an awareness of measurement error is essential here. Henson et al. also provide the following quotes from a review of research practices by Keselman et al. (1998):

One consistent finding of methodological reviews is that a substantial gap often exists between the inferential methods that are recommended in the statistical research literature and those techniques actually adopted by applied researchers (p. 351).

Substantive researchers need to wake up both to the (inappropriate) statistical techniques that are currently being used in practice and to the (more appropriate) ones that should be being used. Quantitative researchers need to wake up to the needs of substantive researchers. If the best statistical developments and recommendations are to be incorporated into practice, it is critical that quantitative researchers broaden their dissemination base and publish their findings in applied journals in a fashion that is readily understandable to the applied researcher (p. 380).
One could also add that it is crucial for educators to include relevant concerns, such as the impact of measurement error, into the curriculum, as a more effective way to disseminate current best practices.

Looking towards statistical thinking, one implication may be the need for greater communication of the problems. In robust statistics, for example, there is a well-developed literature regarding methods to handle outliers (e.g., Maronna et al., 2006), but there is less emphasis on presenting these methods and the need for them to applied researchers, although there are some recent exceptions (e.g., Lind & Zumbo, 1993; Liu, Wu, & Zumbo, 2010; Liu & Zumbo, 2007, 2012; Liu, Zumbo, & Wu, 2012). Although a systematic treatment of the inflation in Type I errors that can result from measurement error was only presented in 2009, many statisticians and methodologists seem to be aware of this and other problems that measurement error can cause. Yet there remains a need for greater communication of these problems to applied researchers. Towards this end, a schematic diagram that may help in organizing some of the subtle problems that measurement error can cause for statistical inference in regression is presented below in Figure 9. This figure is a simplified version of the Draper-Lindeley-de Finetti (DLD) framework introduced in Zumbo (2007) and related to the work of Draper (1995). The explanation following the figure expands and elaborates on points regarding Type I error rates made above, and by DeMars (2010) and Chang, Mazzeo and Roussos (1996).
A1 is the target domain we are ultimately interested in – the relationship among true scores in the population, but C2 is usually the data we have to work with – a sample of observed scores. Traditional methods of hypothesis testing in multiple regression, such as those described in textbooks, provide methods for making inferences from A2 to A1 (path 1) or perhaps C2 to C1 (path 2). Here the statistical model allows for inferences that account for variability due to sampling of people. In the example considered earlier, regarding processing fluency, spatial ability, and mathematics achievement, the data actually collected were C2. When a multiple regression model is used, but ignores measurement error, an inference is essentially made along path 2, to C1. The problem is that C1 ≠ A1 in many cases so that, although we believe we have made an inference along path 3, we have not. Using the earlier notation, making an inference about A1 based on C2 (path 3) is equivalent to basing a decision about $H_0: \beta_x = 0$ on the statistical test of $H_0: \beta_x^* = 0$. One way to think about measurement error models is that they move from C2 to A1 via paths 4 and 1. That is, both the measurement error and sampling error are taken into account in an effort to make the inferences of interest.

---

**Figure 9. Schematic of Inferences in Regression.**
5.6 Novel Contributions

While the problem of inflated Type I error rates for multiple regression was already known in the statistical literature (e.g., Brunner & Austin, 2009), and has been mentioned in related cases in the methodological literature (e.g., Kahneman, 1965; Lord, 1960; Zinbarg et al., 2010), there is not widespread appreciation for the problem. This is evidenced by the common use of hypothesis tests in multiple regression, even when predictor variables are measured with error. The primary contribution of this thesis is providing a clear and explicit statement that measurement error can inflate Type I error rates in multiple regression, and providing an explanation of the problem that is accessible to an audience beyond statisticians. This was done by demonstrating the problem with simulations, explaining why the problem occurs, and discussing which factors were most important in causing it. This thesis also suggested some implications for the teaching of statistics, including specific concepts that could be introduced, and a way to use the DLD framework to assist in conveying the problem, neither of which appear to be addressed in prior studies documenting the issue.

From a more technical standpoint, this thesis also highlighted the distinction between standardized and unstandardized regression coefficients, pointing out the differing ways these parameters are impacted by measurement error. Prior studies either addressed standardized coefficients (e.g., Zinbarg et al., 2010) or unstandardized coefficients (e.g., Gustafson, 2004), without discussion of the distinctions. This thesis also contributed further empirical evidence that Type I error rates are a serious problem even in scenarios that are likely to be encountered in practice, which include cases with smaller effect sizes and additional measurement error in $y$. Finally, the results of the
simulations in Chapter 4 were analyzed and interpreted using the Pratt index, which appears to be a novel application of both the Pratt index and the generalized Pratt index.

### 5.7 Limitations and Future Directions

There are a few relevant qualifications to the findings in this thesis. First, as with any simulation study, one needs to be cautious about making inferences beyond the conditions included in the simulation. The random factors design in Parts I and II of Study 2 was used to support inferences across a wider space of possible conditions, but cannot support other generalizations. For example, all variables were normally distributed. While this is a common scenario to study, and is an assumption of many parametric statistical models (including multiple regression),\(^\text{17}\) the results may differ for other types of distributions. Brunner and Austin (2009) found similar results across three additional distributions for \(\tau_x\), \(\tau_u\), and the measurement errors, which supports the generalizability of the findings, but still leaves the question open. In addition, all measurement errors were additive, non-differential, and unbiased, so claims about more complex measurement error structures are not necessarily warranted.

Another potential limitation involves the method used to manipulate reliability. In order to manipulate the reliability of the different variables, true score variances were held constant while error variance was either increased or decreased. The problem is that Type I error rates (or power) are not a function of reliability, although they are often related (Zimmerman, Williams, & Zumbo, 1993). The Type I error rate will depend directly on the observed score variances and covariances; different ways of manipulating

---

\(^{17}\) Technically the assumption is that the regression error is normally distributed, which often has the effect of leading to a normally distributed outcome variable. The predictors need not be normally distributed to allow for valid inferences.
the reliability will impact the observed score variances differently, and may change
conclusions (Williams, Zimmerman, & Zumbo, 1995). Alternative ways to decrease (or
increase) the reliability include holding the error variance fixed but increasing (or
decreasing) true score variance, or holding the total observed score variance fixed and
simultaneously changing the proportion of true and error score variance. The method here
was selected because it meant that across different levels of reliability, the underlying
true score model remained invariant. Hence, this study addressed the more specific
question, *if the underlying true score relationships remain the same, how are Type I error
rates impacted as measurement error variances changes (i.e., as reliability changes)?*
This version would seem to have the most relevance in applied settings.

It should also be noted that the impact of measurement error on statistical power
(i.e., hypothesis testing when the null hypothesis is false) is a separate question not
addressed here. It could be that in other cases measurement error does lead to a reduction
in power, as is commonly believed. For example, the hypothesis test of $\beta_u$ in the above
simulations was not studied, and could be a topic for future research. However, it is not
meaningful to interpret a hypothesis test and evaluate its power if Type I error rates are
incorrect (e.g., Zumbo & Jennings, 2002), so the present study suggests some limitations
on the conditions in which it would make sense to study that question in the future.

Important directions for future studies include describing and documenting how
more complex error structures can impact analyses, as well as discussing when those
errors are most likely to occur. Extensions to more complex regression models, additional
multivariate techniques, and scenarios where an incorrect model is used would also be
beneficial. These future studies would be important because a) the results here suggest
the impacts could be even more complex and problematic in those situations, and b) as more advanced techniques become widely available in software packages, it is important for researchers to be informed of potential problems. Finally, because in many cases researchers may not have adequate data to fit structural equation or other latent variable models, another essential direction for future research includes developing and evaluating measurement error models that can be used with the limited data that are available to researchers. This research would include dissemination of these methods, so that they make their way into applied practice.

The aim of the simulations in this thesis was not primarily about generating new knowledge – it was already apparent that Type I error rates would be inflated. Rather, the purpose of the simulations was pedagogical and rhetorical – to assist in communicating a complex methodological phenomenon to a wide audience of applied researchers, and to determine the kinds of situations most likely to be affected. Hence, although Parts I and II of Chapter 4 were the most useful from a theoretical and methodological standpoint, a simulation along the lines of Part III, or perhaps an extension of Part III covering additional conditions, appears to have the greatest potential as a rhetorical device for inclusion in a manuscript prepared for publication.
Bibliography


Burks, B. S. (1926). On the inadequacy of the partial and multiple correlation technique. *Journal of Educational Psychology, 17*(8), 532.


Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124


validity: *Revisions, new directions and applications* (pp. 65–82). Information Age Publishing.

