ABSTRACT

Title of Dissertation: INVESTIGATING UNCERTAINTY WITH FUNGIBLE PARAMETER ESTIMATE ANALYSIS

Jordan Yee Prendez
Doctor of Philosophy, 2020

Dissertation Directed by: Professor Jeffrey R. Harring
Department of Human Development
and Quantitative Methodology

Researchers need methods for evaluating whether statistical results are worthy of interpretation. Likelihood functions contain large amounts of information regarding the support for differing estimates. However, maximum likelihood estimates (MLE) are typically the only set of estimates interpreted. Previous research has indicated that these alternative estimates can often be computed and represent data approximately as well as their MLE counterparts. The close fit between these alternative estimates are said to make them fungible. While similar in fit, fungible estimates are in some cases different enough (from the MLE) that they would support alternative substantive interpretations of the data. By calculating fungible parameter estimates (FPEs) one can either strengthen or weaken one's inference by exploring the degree in which diverging estimates are supported. This dissertation has two contributions. First, it proposes a new method for generating FPEs under a broader definition of what should constitute fungible parameter estimates. This method allows for flexible computation of FPEs. Second, this method allows for an exploration of research inquiries that have been largely unexplored. What are the

circumstances in which FPEs would convey uncertainty in the parameter estimates? That is, what are the causes of uncertainty that are measured by FPEs. Understanding the causes of this uncertainty are important for utilizing FPEs in practice. This dissertation uses a simulation study in order to investigate several factors that might be encountered in applied data analytic scenarios and affect the range of fungible parameter estimates including model misfit. The results of this study indicate the importance of interactions when examining FPEs. For some conditions, FPE ranges indicate that there was less uncertainty when the model was correctly specified. Under alternative conditions, FPE ranges suggest greater uncertainty for the correctly specified model. This example is mirrored in several results that suggest that a simple prediction of the level of uncertainty is difficult for likelihoods characterizing real world modeling scenarios.

INVESTIGATING UNCERTAINTY WITH FUNGIBLE PARAMETER ESTIMATE ANALYSIS

by

Jordan Yee Prendez

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Jeffrey R. Harring, Chair
Professor Gregory R. Hancock
Professor Xin He
Professor Hong Jiao
Professor Ji Seung Yang

Acknowledgments

I am exceedingly grateful to my advisor, Dr. Harring, for providing feedback and guidance on this project. I especially appreciated your availability and willingness to meet frequently with me. Like many of your students, I am thankful for your advice but also for your good cheer, which made what could be a difficult time into a positive one. I would also like to thank my dissertation committee for the detailed feedback which greatly improved this work. Thanks to my classmates in the 2017 SEM seminar for the stimulating discussion that piqued my interest in fungible parameter estimation and led me to my eventual dissertation topic. I would like to thank Dr. Hancock as well for meeting with me at the AERA conference in Philadelphia to discuss my interest in the EDMS program. I very much appreciate the positive environment that you and the EDMS faculty have built, where students can thrive in pursuing a variety of interests and career paths. Also, I would like to show my deepest gratitude to my wife, Jennifer, for all those countless hours that you supported (e.g., fed me tasty food) and put up with me as I worked on this dissertation and PhD. Thank you to my family who likewise always backed me and without them I could not have completed this project.

Table of Contents

Acknowledgements ii

Table of Contents iii

List of Tables vi

List of Figures vii

1 Introduction 1
- 1.1 A Statistical and Methodological Challenge . . . 1
- 1.2 Post-Normal Science . . . 2
- 1.3 Fungible Parameter Estimates . . . 3
- 1.4 This Dissertation's Contribution . . . 5
- 1.5 Outline of the Remaining Chapters . . . 7

2 Literature Review 10
- 2.1 Classes of Evidence . . . 11
- 2.2 Sensitivity and Uncertainty Analysis . . . 13
  - 2.2.1 Sensitivity Analysis . . . 13
  - 2.2.2 Uncertainty Analysis . . . 13
  - 2.2.3 Local Methods . . . 15
    - 2.2.3.1 Derivative-Based Local Approaches . . . 15
  - 2.2.4 Global Sensitivity Analysis . . . 16
    - 2.2.4.1 One Factor At a Time Approaches . . . 16
    - 2.2.4.2 Elementary Effects . . . 17
    - 2.2.4.3 Regression Based Sensitivity Analysis . . . 18
    - 2.2.4.4 Generalized Likelihood Uncertainty Estimation . . . 20
- 2.3 Sensitivity Analysis and Related Techniques in SEM . . . 23
  - 2.3.1 Case Diagnostics . . . 23
  - 2.3.2 Model-Specification . . . 24
- 2.4 Model fit . . . 25
  - 2.4.1 Model Discrepancy Function . . . 26
    - 2.4.1.1 Chi-Square . . . 27
    - 2.4.1.2 Absolute Model Fit in FPE Analysis . . . 27
  - 2.4.2 Parsimony Fit Indices . . . 29
  - 2.4.3 Incremental Fit Indices . . . 30

2.4.4 Akaike Information Criterion (AIC) and Related Measures 30
2.4.4.1 Reviewing AIC Interpretation 32
2.4.4.2 Bayesian Information Criteria 33
2.5 Model Indeterminacy in SEM 35
2.5.1 Equivalent Models 35
2.5.2 Causal Search 35
2.6 Fungible Parameter Estimates 37
2.6.1 Observed Variables 37
2.6.2 SEM Generalization 39
2.6.2.1 Multi-Dimensional Approaches 41
2.6.3 Differences With Other Measures of Uncertainty 42
2.6.4 Theoretical Justification for FPE Analysis 43
2.6.5 Generalized FPE Analysis and the PSINDEX 47
2.6.5.1 Parameter Stability Index 48
2.6.5.2 Simulated Annealing Method 49
2.6.5.3 Differences With the Existing FPE Methodology 50
2.6.5.4 Guidelines for Use and Interpretation 51

3 Methods 60
3.1 Research Questions 61
3.2 Simulation Design 62
3.2.1 Measuring Uncertainty – Outcome Measures 62
3.2.2 The Utility of Monte Carlo Simulations for Studying FPE Analysis . 63
3.2.3 Data Generation 67
3.2.3.1 Procedure for Non-Convergence 68
3.2.4 Assessing Fidelity and Variance of Data Generation procedure 69
3.2.4.1 Simulated Population-Level Data Fidelity 71
3.2.4.2 Sample Variability and Monte Carlo Error 71
3.2.5 Replications and Simulated Annealing Maximum Iterations 76
3.3 Manipulated Factors 77
3.3.1 Estimated Models 77
3.3.2 Model Fit Index 80
3.3.2.1 RMSEA as a Measure of Model Decrement 80
3.3.2.2 AIC as a Measure of Model Decrement 82
3.3.3 Sample Size 83
3.3.4 Model Complexity 84
3.3.5 Measurement Quality 85

4 Results 88
4.1 Data Evaluation 89
4.1.1 MCE ML Estimates 89
4.1.2 MCE FPE Percent Change Metric 90
4.1.3 MCE FPE Range Metric 91
4.2 Outcome by Level of Summary 92
4.2.1 Overall Results 92

4.2.1.1 Exploring Results with Eta-Squared . . . 92
4.2.1.2 RMSEA and AIC . . . 93
4.2.1.3 RMSEA . . . 95
4.2.1.4 AIC . . . 97
4.2.2 Results by Parameter Type Subgroup . . . 97
4.2.3 Summary of Overall and Parameter Type Subgroup Results . . . 101
4.2.4 Results by Individual Parameter . . . 102
4.2.4.1 Sample Size . . . 102
4.2.4.2 Model Complexity . . . 103
4.2.4.3 Model Complexity*Fit Decrement (RMSEA) . . . 105
4.2.4.4 Model Complexity*Sample Size (AIC) . . . 108
4.2.4.5 Model Condition . . . 110
4.2.4.6 Model Condition*Fit Decrement (RMSEA) . . . 115
4.2.4.7 Measurement Quality . . . 116
4.2.4.8 Effect of Model Misspecification, Measurement Quality, and FPE Index on Parameters of Interest . . . 117
4.2.4.9 FPE Range and FPE Global Range . . . 120
4.2.5 Summary of Study Factors . . . 121
4.2.5.1 Sample Size . . . 124
4.2.5.2 Model Complexity . . . 125
4.2.5.3 Model Condition . . . 125
4.2.5.4 Measurement Quality . . . 126
4.2.5.5 FPE Index and Fit Decrement . . . 127

5 Discussion 133
5.1 Summary of Results and Implications . . . 134
5.1.1 Testing the FPE Framework . . . 139
5.1.2 Extensions and Contributions to Research . . . 141
5.2 Study Limitations and Future Research . . . 145
5.3 Conclusion . . . 147

A Methods Section Pilot Study Results 149

B Methods Section – Auxiliary Tables and Figures 153

References 162

List of Tables

3.1 Simulated Data Fidelity 70
3.2 Monte Carlo Error by Replication and Max. SA Iteration 77
3.3 Average Number of FPEs Estimated and Study Time 77
3.4 FPE Percent Change Estimate by Replication and Max. SA Iteration 78
3.5 Proposed Manipulated Factors. 87

4.1 $\hat{\eta}^2$ for Factors Representing the Manipulated Conditions 94
4.2 $\hat{\eta}^2$ Factors Representing the Manipulated Conditions (RMSEA) 96
4.3 $\hat{\eta}^2$ Factors Representing the Manipulated Conditions (AIC) 98
4.4 FPE Values for Misspecification 1 and Perfect Conditions by Model Decrement (RMSEA) 122
4.5 FPE Values for Misspecification 1 and Perfect Conditions by Model Decrement (AIC) 123
4.6 FPE Values for Misspecification 2 and Perfect Conditions by Model Decrement (RMSEA) 124
4.7 FPE values for Misspecification 2 and Perfect Conditions by Model Decrement (AIC) 125

List of Figures

2.1 Regression based sensitivity analysis 19
2.2 Peaked log likelihood vs. parameter estimates 45
2.3 Non-peaked log likelihood vs. parameter estimates 46
2.4 Likelihood MLE 52
2.5 Likelihood MLE and FPE contour 53
2.6 Likelihood MLE and FPE contour method 54
2.7 Likelihood MLE and generalized FPE method 55
2.8 Structural Equation Model of the Political Democracy dataset 56
2.9 Political democracy FPE output (1/2) 58
2.10 Political democracy FPE output (2/2) 59

3.1 Example data and likelihood function 65
3.2 Example estimates and likelihood functions 66
3.3 Example likelihood conditions and FPEs 67
3.4 Pilot SEM 72
3.5 Monte Carlo Error 75
3.6 Generating Model 79
3.7 Misspecification Model 1 81
3.8 Misspecification Model 2 82

4.1 Estimated MCE for ML Parameter Estimates 90
4.2 Estimated MCE for the FPE percent change metric 91
4.3 Estimated MCE for the FPE range metric 92
4.4 FPE range by study factor overall 95
4.7 FPE range by variable type (RMSEA) 99
4.8 FPE range by variable type (AIC) 100
4.9 FPE range by parameter (sample size) 103
4.10 FPE range by parameter (model complexity) 105
4.11 FPE range by parameter (model complexity and misspecification) 107
4.12 FPE range by parameter (model complexity and sample size for AIC) 109
4.13 FPE range by parameter (model misspecification with AIC) 110
4.14 FPE range by parameter (model misspecification with RMSEA) 112
4.15 FPE range by parameter (model misspec. 1 and decrement levels with RMSEA) 114
4.16 FPE range by parameter (model misspec. 2 and decrement levels with RMSEA) 116

4.5 FPE range by study factor overall (RMSEA) 129
4.6 FPE range by study factor overall (AIC) 130
4.17 FPE range by parameter (measurement quality and model misspec. 1) . . . 131
4.18 FPE range by parameter (measurement quality and model misspec. 2) . . . 132

Chapter 1: Introduction

1.1 A Statistical and Methodological Challenge

Statistical inference is an important tool that is utilized by researchers as a method for understanding an incredibly wide category of phenomenon. The promise of statistical analysis is that it allows for impartial summaries of data that do not suffer from the same weaknesses of other forms of knowing (e.g., anecdotal evidence). The reality, however, is that the goal of a clear and unbiased truth being portrayed by statistical inference is often very difficult to achieve. This difficulty has become more salient recently with the well publicized finding that several scientific research studies failed to be replicated due to poor statistical technique. One of the most publicized studies was an effort to replicate 100 correlational and experimental studies from three psychology journals. Despite collaborating with many of the original study authors in order to replicate the results, only 39% of the studies were determined to be replicated (Open Science Collaboration, 2015). Using a prediction interval (rather than a binary cutoff) to adjudicate whether replication was statistically significant resulted in 77% of studies determined as replicated (Patil, Peng, & Leek, 2016). A prediction interval is a method of predicting a range in which future values may reasonably fall. Despite this more optimistic replication figure, 80% of those studies judged as replicated had smaller effect sizes than the original study. These types of statistical and methodological problems are not confined to psychology, as other disciplines have also faced criticism that results are not replicable. These smaller effect sizes, and non-replicated effects represent wasted time and money. In medicine, the amount

of wasted resources on non-replicable work was estimated at $28,000,000,000/year (Freedman, Cockburn, & Simcoe, 2015). The types of issues that have caused these failures to replicate are multi-fold. Publication bias, p-hacking, underpowered statistical designs, and outright fraud all contribute.

There is much work to improve statistical methodology and research design (pre-registration, focus on effect size, theory focused design etc.). The challenge of correctly interpreting statistical results is made more difficult in a subset of research: observational studies. Observational studies are those in which the independent variable (e.g., social economic status, race, IQ) is not manipulated by researchers interested in its potential effects on a dependent variable (e.g., income, educational attainment). While experimental design is a powerful tool for elucidating cause and effect, many research questions cannot be reasonably designed as experiments. It may be tempting to avoid observational designs, however, without them researchers would not be able to study some of the most important issues. Many questions regarding education, human behavior, medicine and epidemiology cannot be studied as experiments due to ethical or practical reasons. Unfortunately, observational studies may be even less replicable than experimental studies. Due to the risk of wasted time, money, and other resources, researchers and statisticians must search for methodological tools that can improve our ability to interpret statistical results. A better understanding of this uncertainty would aid researchers in avoiding interpreting statistical results that may not be replicable.

1.2 Post-Normal Science

More broadly, the statistical and methodological issues described above can be thought of as an over-reliance on the belief of science as an infallible method that arrives at unbiased facts that point to particular solutions. Post-normal science (PNS) is a framework that argues that when "facts are uncertain, values in dispute, stakes high and decisions urgent" (Ravetz & Ravetz, 1993, p. 744) that the normal rules of science do not operate. For instance, under "normal science" (Kuhn, 1970), science moves along using a particular

framework for research that scientist agree on (e.g., p-values are useful). Using these rules, the scientific method is used to uncover facts and truth about a particular phenomenon. For instance, in a well-defined area of chemistry, experimentation in conjunction with null hypothesis statistical tests can lead to a causal understanding of the underlying physical processes. Under the conditions for post-normal science listed above (i.e., "facts are uncertain...") the rules of "normal science" do not necessarily apply. A statistical model, for example, might be used to study and determine the effects of factors important to determining whether a small business succeeds or not. In this example there are high-stakes decisions to be made as well as high levels of uncertainty. The purpose of PNS is to draw attention to the high levels of uncertainty that are inherent to many of these high-stakes decisions and to point out that traditional techniques (i.e., p-values, effect size, model fit) are not enough to justify a conclusion. Secondly, PNS also makes salient the value judgements and other factors that are embodied in the results and that interpreting these results as the unbiased arbiters of fact with only one potential solution is a mistake. Instead, under conditions of PNS we should instead aim to acknowledge the uncertainty, biases, and assumptions that can cloud results.

In many ways, questions regarding education, climate science, conservation, and other important areas are considered in the domain of PNS. Under this framework, uncertainty and sensitivity analysis are tools in which one can confront and genuinely address complicated scientific issues in which there are large amounts of uncertainty.

1.3 Fungible Parameter Estimates

One method for assessing how (or whether) statistical results should be interpreted is through the analysis of fungible parameter estimates (FPEs). FPEs are alternative parameter solutions so similar to the maximum likelihood estimates (in terms of data-model fit) that they could be considered fungible (i.e., exchangeable) with the maximum likelihood estimates (MLE). Caution should be exercised when there are large discrepancies between the values of these alternative (FPEs) and the MLEs. FPEs can be

conceptualized in terms of the following expression:

$$(\hat{\theta}_{MLE} \in \hat{\theta}_{FPE} \mid \hat{F}_{\theta_{MLE}} \approx \hat{F}_{\theta_{FPE}}) \tag{1.1}$$

Here, $\hat{\theta}_{MLE}$ and $\hat{\theta}_{FPE}$ are model parameter estimates at the MLE, and slightly less optimized estimates, respectively. The parameters $\hat{F}_{\theta_{MLE}}$ and $\hat{F}_{\theta_{FPE}}$ represent a fit function (generically defined) for the MLEs and at the less optimized points along the function surface, respectively. From the FPE framework, the MLEs are among a larger set of other fungible (i.e., alternative) estimates, given that model fit is substantially equivalent under all estimates.

FPE analysis is an approach in which statisticians or applied researchers can measure the uncertainty of their statistical results. Specifically, analyzing the range of FPEs informs researchers as to how uncertain a particular parameter is in describing the data. Results with high levels of uncertainty should be interpreted as such. For instance, imagine a researcher is interested in the effect of an intervention on a math test. The result of the experiment indicates that those in the intervention group scored higher (i.e., more questions correct) than those not receiving the intervention with the mean effect of being assigned to the treatment group is M = 1.4 (SE = 3.0). While the effect of being assigned to the experimental condition is 1.4, the standard error of 3.0 informs us that there is a large amount of uncertainty regarding the estimate of the treatment effect and we should be wary of interpreting it. This type of uncertainty is commonly assessed and is measured as part of null hypothesis statistical testing (NHST) (Neyman & Pearson, 1933). Analyzing FPEs is an additional method in which a distinct type of uncertainty regarding parameter estimates can be assessed. Using FPEs, an experimenter can assess how stable parameter estimates are, and their level of necessity in describing the data.

Recently, research has demonstrated this second type of uncertainty contained within a latent variable path model. In their model, the MLE for the effect of representation on attachment was $\hat{\theta}_{MLE} = .339$ ($SE = .109$). However, an FPE analysis indicated that an

estimate of $\hat{\theta}_{fpe} = -.350$ was almost equally as likely (T. Lee, MacCallum, & Browne, 2017). Here we can see that there can be a large amount of uncertainty in estimates, despite comparably small standard errors and the accompanying statistical significance. Given that this relatively large range of estimates (i.e., -.350, .601) almost equally well describes the data, many researchers would be cautious of interpreting the maximum likelihood estimate of $\hat{\theta}_{MLE} = .339$. Therefore, FPE analysis can be a powerful tool in assisting researchers in determining whether or not they should interpret statistical results. This type of analysis has come largely out of two different analytical traditions, sensitivity analysis and data model fit. FPE analysis is a combination of these two types of analyses in that it explores how sensitive the statistical data model-fit is to changes in parameter estimates.

1.4 This Dissertation's Contribution

Previous research has indicated that by analyzing the range of FPEs, researchers can gain new insight regarding parameter estimate stability. However, there have been relatively large limitations to the study of FPEs. Estimation limitations restricted the study of FPEs to only a few "focal parameters" and not all model parameter estimates. In addition, this technical limitation has also been accompanied by a somewhat narrow conceptual approach to FPE analysis. Both of these reasons have constrained previous research but have been alleviated, in part, due to an expanded analytic framework and a new R package `psindex` that allows for expanded research on FPE analysis for latent variable models (Prendez & Harring, 2019). FPE analysis is partly a measure of how influential a parameter is in determining overall-fit. Data model-fit is traditionally treated as a omnibus measure of whether the data is well represented by a statistical model. This is undeniably important information, but understanding which parameters, in particular, are worthy of interpretation is additional useful knowledge. With this information, researchers might avoid interpreting individual parameters that do not describe the data,

and contrarily have greater confidence in interpreting those that are important in describing the data. In addition, understanding which parameters are most important in determining the overall model fit may give insight into building statistical models that better represent their respective populations. However, it is unknown to what extent FPE analysis can aid in identifying, and/or remediating local model misfit.

A constrained conceptualization and lack of ability to generate FPEs has limited investigations into the extent to which FPE analysis can assist in parameter interpretation. Under which circumstances do FPE ranges represent cause for concern and when should they be viewed as bolstering the evidentiary support for interpreting parameter estimates? Previous research has shown that FPEs can provide useful information in parameter interpretation but little work has been done to facilitate an understanding of the circumstances in which this type of analysis is valid (e.g., sample size, model complexity, measurement quality, selection of fit-index). FPE analysis has been demonstrated convincingly in the past primarily by small scale real-data examples. There is, however, a dearth of evidence that demonstrates the technique under differing scenarios (previous limitations on calculating FPEs have made these types of studies difficult). It is currently unknown if fungible parameter estimates should be interpreted similarly under only a narrow subset of data analytic circumstances or rather should be interpreted similarly under a broad set of scenarios. Understanding this is critical if FPE analysis is to be used as an analytical tool in model evaluation. Therefore, the contribution of this dissertation is to help elucidate the circumstances in which FPE analysis can be used to assist in parameter interpretation. Based on the results of a Monte Carlo simulation study, suggestions for the interpretability of FPEs under different modeling scenarios (e.g., differing levels of model-complexity, sample-size) are given in addition to suggestions for future research. This is congruent with an ever growing knowledge that hard cut-offs and "golden rules" should be replaced with more contextual guidelines (Heene, Hilbert, Freudenthaler, & Bühner, 2012; Moshagen, 2012).

Furthermore, the purpose of this project is to provide an opportunity to falsify this new methodological technique. For example, imagine the scenario in which an FPE analysis indicates that estimates of -.2 and .5 are both equally valid descriptions of the data. That is, given this information, a researcher would understand that there is a fairly large amount of uncertainty around the parameter estimate and would have less confidence in interpreting this value. However, while this is assumed to be informative, little to no research has been done from a simulation standpoint to assess whether or not this is true. For example, imagine the following scenario in which the usefulness of FPE analysis might be called into question. An FPE analysis of a simulated dataset with perfect fit and large sample size. One such example might include an FPE analysis conducted on a large sample simulated dataset, with a perfectly fitting model. The FPE analysis then reveals FPEs of -.2 and .6 for a parameter with an MLE of .3. This is problematic because the FPE analysis is pointing to a problem when none exists. The purpose of this dissertation is to elicit the cases in which FPE analysis is useful and the cases in which it might lead researchers to incorrect conclusions. In short, FPE analysis is thought to give researchers certain information leading to particular interpretations. This dissertation aims to test whether FPE analysis, in its current form, can be falsified as a methodology. The ability for a theory to be falsified, is a cornerstone of the scientific method (Popper, 1959). Without having an understanding of the causes of large or small fungible ranges, FPE analysis risks being a tool reported only when it is in alignment with a researchers preferred hypothesis and ignored otherwise. This risk is exacerbated if the ranges are to a large degree random rather than being related to the uncertainty of the parameters or another useful quality in which useful inferences might be made. Thus the theory tested is: Do FPEs encode information – via the likelihood ratio – regarding whether a particular parameter estimate should be interpreted?

1.5 Outline of the Remaining Chapters

In summary, Chapter 1 presented a few examples demonstrating the need for statistical and methodological improvements. This need is perhaps more acute in studies in which randomized control trials are not feasible. FPE analysis was described as an analytical tool for either strengthening or weakening one's inferences about individual parameter estimates. Because the type of uncertainty measured by FPE analysis can be characterized as a description of how well individual parameters describe the data (i.e., data model fit), it may also represent a tool in identifying and remediating model misfit. The goal of this dissertation is to examine the circumstances in which FPE analysis is valid, and how FPEs may be interpreted in differing scenarios.

Chapter 2 brings together relevant literatures beginning with an overview of sensitivity analysis research, followed by a review of the model-fit literature. Next, a review of current FPE research is presented. Justification as to why FPEs might be interpreted as well as the introduction for a new generalized FPE framework is also given.

Chapter 3, the methods section, reviews the design of a simulation that is used to investigate the relation between FPEs ability to quantify uncertainty under a variety of data analytic conditions found in practice. This section describes the data-generation technique and the factors manipulated within the study. Manipulated variables and study conditions are justified in terms of why they might be expected to affect the ability of an analysis of FPEs to quantify uncertainty. Included in this section is also justification of why the levels chosen represent reasonable choices for investigation. Finally, a description of the outcome measure is operationalized.

Chapter 4 presents the outcomes of the simulation study. These results are preceded by a summary of the precision of the simulation study. The results are then presented at three levels of summary beginning with the most general. The first, and highest level of analysis presented is the overall results section which are summarized across all model parameters. The second set of results are summarized across parameter type. Finally, the most granular set of results are presented by individual parameter. This section is concluded by a more

summative analysis describing the results by study factor.

Chapter 5 begins with an overview of the simulation results and the most important findings. This section interprets these more surprising results in addition to considering the overall pattern of outcomes. The meaning of these results are also discussed in reference to their role in evaluating the FPE methodology, and in relation to previous simulation work. This chapter concludes with a discussion of the ways in which the current work was limited and suggestions for future research on FPE analysis.

Chapter 2: Literature Review

Statistical models are methods in which researchers apply mathematical descriptions to natural phenomena in order to learn about, or predict a given system. Statistical models are almost always simplifications of the phenomenon they are modeled after, and are thus only approximations. Models must then be evaluated to help the user understand the degree to which these approximations are representative and useful. There are many types of assumptions that a researcher must make in order to conduct an analysis. These different choices all have the potential of influencing the results and validity of the conclusions drawn. FPE analysis involves exploring, and making salient, the uncertainty present in results.

FPE analysis can be traced to several related topics, including many under the more general topics of sensitivity analysis (SA) and uncertainty analysis (UA). Research regarding fit and equivalent model research is also closely related to FPE analysis and will also be reviewed. This chapter will also review maximum likelihood estimation from within a structural equation modeling framework and how it relates to FPE analysis. Finally, justification for using FPEs as a method for model evaluation will also be given in terms of likelihood theory. The purpose of this review is two-fold: First, to understand how FPE analysis relates to and differs from other similar techniques. Those who have confidence in similar approaches will, with hope, view FPE analysis as a logical extension of related techniques. Secondly, in order to implement a meaningful test of the FPE analysis framework, a review of the underlying theory and current research is needed.

2.1 Classes of Evidence

To best interpret FPEs, it is useful to understand how they can be placed within the larger statistical landscape. Most generally, statistical modeling can be broken down into purposes: those used primarily for causal explanation (e.g., does smoking cigarettes cause cancer?) or those used primarily for prediction (e.g., Netflix's recommendation engine, Google's photo recognition; Shmueli, 2010). These goals, of course, are not mutually exclusive. A misunderstanding of the purpose of the research question actually addressed (i.e., causal or predictive) versus the intended one is one of the most common problems in statistical analysis (Leek & Peng, 2015). In both approaches, results must be evaluated in terms of whether ensuing results shall be considered valid for the purpose in which the analysis was designed.

For statistical modeling in which prediction is the goal, models are evaluated primarily in terms of their predictive power. However, in many academic disciplines (e.g., Education, Psychology, Epidemiology, Medicine), the primary goal is an understanding of the causal mechanisms underlying the phenomena of interest even when only observational methods are available. It is within this framework that statisticians and other researchers use multiple sources of evidence (e.g., model fit, likelihood, probability, fungible parameters estimates) to build a case for —or at least maintain the possibility of —a particular causal explanation.

Statistical models can be evaluated using many different sources of evidence and an analysis of fungible parameter estimates (FPEs) provides researchers an additional tool to evaluate them. This method will be compared to other approaches used in the model evaluation process. The purpose is to both differentiate FPEs from other types of model evaluation while also demonstrating ways in which an examination of FPEs is consistent with the underlying logic used by other types of model evaluation strategies on which researchers already rely.

Traditionally, statistical models start with an evaluation of whether they meet a set of assumptions that depend on the type of statistical procedure (e.g., linearity assumption in Ordinary Least Squares (OLS) regression, satisfaction of the multivariate normality when assumed in maximum likelihood estimation, etc.). While meeting the assumptions of the particular technique is considered as important, they are not viewed as imbuing the results with any sort of significance. However, checking assumptions is an important early stage in the analytic pipeline. There can be no meaningful analysis of results before verifying the assumptions of which those same results are based. Traditional model-fit indices and an examination of FPEs should also be viewed in a similar light.

While examining whether or not particular coefficients (e.g., regression slope, latent variable path, etc.) are statistically significant is an important part of the analysis and inference process, this dissertation is interested in factors that are used prior to the interpretation of statistical significance and standard errors.

Several related approaches to analyzing fungible parameter estimates will be reviewed. In particular, model fit (including related equivalent equivalent model analysis, and causal search algorithms) and sensitivity analysis are discussed because they are similar methods for gathering evidence that can either strengthen or weaken the plausibility of model results. Model fit as discussed is a technique for linear-SEM, whereas causal search is a technique that can be instantiated for parametric or non-parametric models. It is important to note that while some of these techniques are extremely varied from a technical perspective they share elements of a common philosophical approach. That is, proving a statistical result is often difficult (or impossible) but evidence can be gathered in order to evaluate the plausibility of a model or result. FPE analysis is another methodology in line with these types of approaches rather than a strict binary testing framework (e.g., null-hypothesis statistics testing).

2.2 Sensitivity and Uncertainty Analysis

2.2.1 Sensitivity Analysis

Sensitivity analysis is a set of techniques that is characterized as measuring the potential impact of any factor that has an effect on some output variable of the model (e.g., modeling choices, data sub-setting, distributional assumptions etc.) (Saltelli et al., 2008). Because FPE analysis can be characterized as a type of sensitivity analysis, it is important to understand the relevant techniques and history associated with this methodology.

2.2.2 Uncertainty Analysis

Before a sensitivity analysis can take place an analysis of uncertainty must precede it. Uncertainty analysis, and sensitivity analysis are related techniques. In an uncertainty analysis, there is a quantification of the how uncertainty in inputs affects outputs. Whereas, in a sensitivity analysis the researcher attempts to determine how the uncertainty in the inputs can be allocated to different factors (i.e., which factors are most important). These two analyses are often times conducted in tandem. In order to explain how sensitivity analysis is conducted, a broad overview is first given with further details on the steps to follow.

Before apportioning uncertainty in model output to particular factors, it is first necessary to determine how uncertainty in input factors propagates to outcome uncertainty (i.e., uncertainty analysis). The steps needed to conduct an uncertainty analysis are now detailed. First, a determination of the set of factors $(X_1, X_2, ..., X_n)$ that is to be included in the analysis must be undertaken. Second, the distributions $(D_{X_1}, D_{X_2}, ...D_{X_n})$ for each factor is set. Third, a sample of y outcomes is generated based on the uncertainty range as specified in the distributions of the input factors. Once a set of y outcomes has been generated by varying the input factors and estimating the model, a sensitivity analysis can take place by apportioning which factor is most responsible for the changes in the outcome variable y.

In order to conduct an uncertainty analysis, however, care must be taken with each of

the steps listed above. The decision of which factors should be included is a non-trivial step. Omitting too many factors can undermine the validity of an uncertainty analysis as it is not possible to quantify the contribution of factor X_1 on outcome Y if outcome X_1 is not included in the analysis. In contrast, including too many factors can greatly increase the computational costs and make the overall analysis unfeasible. Importantly, however, Saltelli et al. (2008) suggests that it is often the case that a few variables are overwhelmingly responsible for uncertainty in the output, while the remainder of variables contribute to a relatively minor degree. As Saltelli et al. (2008) pointed out, this is helpful because once the primary factors driving the uncertainty have been added, adding others will not greatly increase the uncertainty in the output but can also help to satisfy potential critics who believe a particular variable should be included.

Several methods exist for determining the range of factors to include; this process is often time consuming Saltelli (2004, 2008a). Because these values represent the inputs to the given uncertainty analysis they must, as best as possible, represent the uncertainty in the inputs. Methods of generating the distribution may be based on a review of the literature, or on a normal distribution when utilizing values from a literature review is impractical or not possible. In either case, the values chosen must be defended. If these values are not viewed as representing reasonable potential values by critics, the resulting sensitivity analysis will likely be less meaningful.

The purpose of sensitivity analysis is to aid researchers in evaluating whether statistical results should be interpreted. Slight changes to the input resulting in correspondingly large changes in output can signal an underlying instability, or misspecification, and cautions researchers from over-interpreting a particular result. Conversely, when results are robust, researchers may have more confidence in their subsequent interpretations. While it is generally not possible to prove a model, a sensitivity analysis can reveal the extent to which a model might be falsified and how unstable the outputs are in regards to changes in input values.

2.2.3 Local Methods

In general, sensitivity analysis and uncertainty analysis are categorized under one of two categories, local or global. To give context to these two methods it is beneficial to consider the mechanism under which parameters are estimated. To estimate parameters, in a non-Bayesian context, some form of mathematical optimization must take place. Ideally, this optimization results in a clear optimum point that generates an ideal set of values that best satisfies the given function (e.g., maximum likelihood estimates from maximizing a likelihood function, or OLS estimates from minimizing the least squares criterion function). This optimized solution serves as the point in which local, partial-derivative tests are based. Below is a description of this derivative-based approach to uncertainty and sensitivity analysis.

2.2.3.1 Derivative-Based Local Approaches. Using this type of method, the rate of change in Y is measured by taking the partial derivative of the dependent variable with respect to each factor X_i. The general formula for this type of approach is shown below in Equation (2.1).

$$S_i = \frac{\partial Y}{\partial X_i}. \tag{2.1}$$

Here, X_i represents each factor in the sensitivity analysis, Y is the output of interest and S_i is an index of sensitivity. This coefficient can then be used to rank the subsequent variables in terms of their impact on Y. Therefore, the sensitivity in Y can be apportioned to the variables with the largest changes in Y. This same approach can be generalized to circumstances in which we are interested in the interaction between two variables, and more complex relations, by using second, third and nth order partial derivatives (Razavi & Gupta, 2015). For instance, in order to investigate second-order interactions between variables, a calculation of sensitivity coefficients based on the second-partial derivative is appropriate:

$$S_{ij} = \frac{\partial^2 Y}{\partial X_i \partial X_j}. \tag{2.2}$$

From a frequentist approach, statistical inference based on maximum likelihood estimation is concentrated at the maximum of the likelihood function. The partial derivative approach, therefore, has the advantage of being intuitive in that the uncertainty analysis is constrained to investigating the properties of this estimate. However, this type of analysis describes only a small section of the likelihood function. In cases in which the model is simple, a local sensitivity analysis should generalize well to the rest of the input space (Razavi & Gupta, 2015; Saltelli, 2008a, p. 11).

This derivative-based method can be improved by adding two vectors describing the variability (i.e., uncertainty) in the inputs and outputs. Given certain conditions, this methodology can produce similar results to other global based techniques discussed below (Saltelli et al., 2008).

$$S_i^{\sigma} = \frac{\sigma_{X_i} \partial^2 Y}{\sigma_y \partial X_i}.$$

Here, σ_Y represents the variability of the output and σ_{X_i} represents the variability in the input X factors. Whereas, the distribution of the input factors (D_{X_i}) is set at a range determined to be defensible by the researcher.

2.2.4 Global Sensitivity Analysis

Compared to local sensitivity analysis, global uncertainty and sensitivity analysis attempts to quantify and apportion uncertainty over the entire input factor space rather than at one optimal point as in local sensitivity analysis. Several of these global methods are now discussed.

2.2.4.1 One Factor At a Time Approaches. One factor at a time (OAT) approaches explore the input factor space by varying each individual factor independently. OAT approaches are the most popular method within global sensitivity analysis (GSA)

methods (Saltelli & Annoni, 2010). To conduct an UA / SA using OAT methods, researchers must begin with steps similar to other methods. First, a set of factors to be included in the UA/SA analysis (e.g. which factors should vary?) must be determined. Once these variables are chosen, distributions are set for each of these factors. Next, the range of possible Y_i outputs is generated by varying each factor (i.e., model parameter, observations included, model constraints) across the distribution specified. This is done while holding each other variable constant (e.g., often at the Maximum likelihood estimate for a given parameter). This process is repeated for one factor at a time until each factor has been varied and the resulting distribution of Y_i has been generated.

This approach is useful in that it considers a larger proportion of the input space that are far from the baseline (i.e., often the MLE) and acknowledges that local approaches may be insufficient for describing more complicated functions. In addition, because only one factor is being modified at a time, any change in the output can always be localized to that individual input factor (Saltelli & Annoni, 2010). However, there is reason to believe that this approach does not represent the input factor space very well. Saltelli and Annoni (2010) indicate through a geometric proof that as the number of input factors increases, the proportion of the input factor space that is explored by OAT approaches decreases. In an example, the authors demonstrate that the proportion of the input factor space that can be sampled using OAT methods for a model with 2, 3, and 12 factors is .77, .52, and 0.000326, respectively. The very low percentage of the target uncertain factor space accessible by OAT methods demonstrates that it is not appropriate for assessing the input surface especially for models containing more than a few factors.

2.2.4.2 Elementary Effects. Morris's elementary effects method can be seen as a modified version of OAT methods aimed at addressing the weakness of OAT methods by better representing the target input factor space (Morris, 1991; Razavi & Gupta, 2015). The Elementary Effect's method is defined as:

$$EE_i = \frac{y(x_1, x_2, ...x_{i-1}, x_i + \Delta x_{i+1}, ..., x_k) - y(x_1, ..., x_k)}{\Delta}, \tag{2.3}$$

Here, k represents the number of independent factors. Each one of these factors is defined to have p levels. The formula for calculating a sensitivity index under this method is shown below in Equation (2.4).

$$\mu_i^* = \frac{1}{r} \sum_{j=1}^{r} |EE_i^j|. \tag{2.4}$$

Here, r distinct starting points are utilized in order to obtain an average measure of parameter influence, where μ_i^* represents the overall average parameter influence for the ith input factor. Rather than using one point (e.g., the optimized parameter values), the elementary effects method determines the sensitivity of each input factor averaged over r points (Campolongo, Cariboni, & Saltelli, 2007; Morris, 1991).

Elementary effects are viewed as a suitable approximation of more computationally intense methods (Pianosi et al., 2016; Razavi & Gupta, 2015). The following methods are a sampling of the types of methods used to conduct global uncertainty analysis (GUA) and or global sensitivity analysis (GSA).

2.2.4.3 Regression Based Sensitivity Analysis. Regression based sensitivity analysis is a method for GSA that allows for an interpretable approach for summarizing and apportioning uncertainty. In order to conduct this type of analysis, an investigator must again begin with a determination of the factors to include in the analysis, and their respective distributions. Once this step is completed, a comprehensive number of model instantiations is undertaken in order to synthesize data under a large, and ideally representative, set of study parameter combinations. The following example illustrates the general procedure.

To illustrate a regression based GSA, consider a model with three input variables X_1, X_2, and X_3 that are thought to influence a single output variable Y. There is an unknown

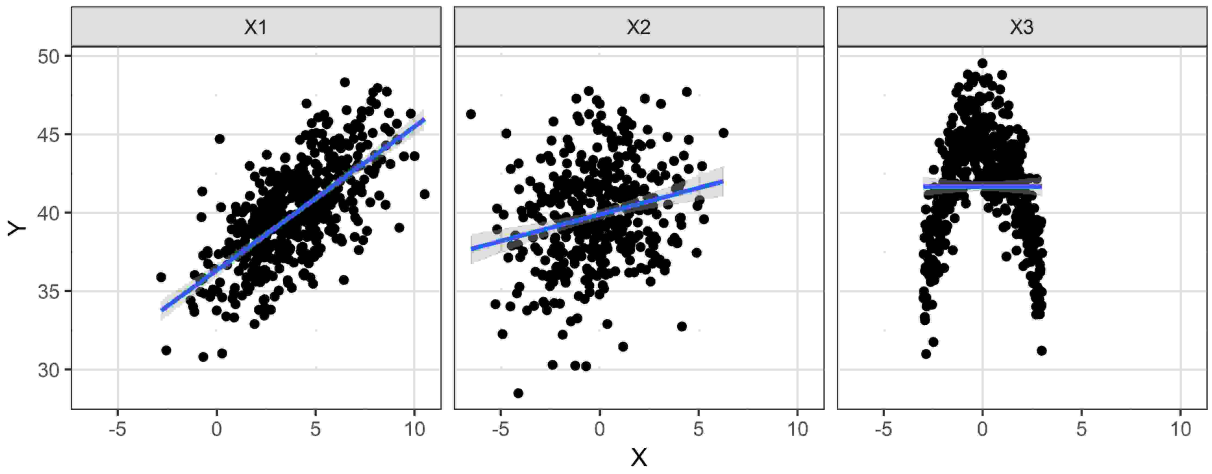

Figure 2.1. Scatter plots of the three input factors and their effect on the outcome variable Y. The line of linear best fit is shown in blue. Each point represents a represents a unique model solution that is calculated by varying the respective input variable (i.e., uncertainty analysis results).

level of uncertainty in each of the three variables. After careful consideration, the researcher assigns distributions representing the uncertainty for each input factor as $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{U}(0,1)$, and $X_3 \sim \mathcal{N}(0,1)$.

The uncertainty is then quantified by generating model instantiations representing a sample of the three input distributions and recording the resulting output y_i. Each value of y_i is then taken as the data points in which a regression analysis is then used to quantify the relative sensitivity of the respective input variable on the Y output. Because bivariate scatterplots form a familiar method of assessing the strength of linear relationship, this method serves as an intuitive method for sensitivity analysis.

Similar to the examples by Saltelli (2008b), Figure 2.1 demonstrate theoretical data points and the regression sensitivity analysis technique. Data would be generated as part of the uncertainty analysis. Each point represents a model instantiation within the input factor space (i.e., each point represents a single result from the uncertainty analysis). By examining the strength of their linear relationship, the sensitivity of the output to input factors can be quantified. As can be seen in Figure 2.1, X_1 is a very important factor in determining the value of Y whereas X_2 is a less important factor.

The data in the scatterplots serve as more than a tool for informally assessing sensitivity. These data, as the method suggests, are utilized in a linear regression analysis in order to quantify the strength (i.e., β_{x_1}) of each relationship for each predictor in order to rank and quantify Y output sensitivity to uncertainty in predictors X_i. Because this model uses a linear regression in order to rank the sensitivity of Y by predictor X_i it becomes less useful as a method when the relationship between Y and the predictors X become less linear (Saltelli, 2008b). This nonlinear scenario can be seen in the rightmost panel in Figure 2.1. Here Y is highly sensitive to X_3 but a beta coefficient representing this nonlinear relationship would serve as a very poor descriptor of this strong sensitivity.

2.2.4.4 Generalized Likelihood Uncertainty Estimation. Generalized Likelihood Uncertainty Estimation (GLUE) is another global sensitivity analysis method. This methodology differs from previously mentioned methods in that it ranks models in terms of their likelihood performance. This method is also similar to FPE analysis in that it only considers values that are probable based on their value of the likelihood. This method was invented for use originally in evaluating complicated hydrological models (Beven & Binley, 1992). The GLUE methodology followed upon results from Monte Carlo sensitivity analysis of hydrological models indicated that a multitude of hydrological models provided similar predictions (e.g., as measured by Nash-Sutcliffe index values Nash & Sutcliffe, 1970) rather than one global optimum model. At the time, this new framework, was created in order to highlight and measure this type of uncertainty. In order to conduct a GLUE analysis, researchers first classify models as either "Behavioral" or "Non-behavioral models". Models deemed "non-behavioral" are not considered further within the GLUE framework. For instance, model parameters leading to output that violates a physical law (i.e., liquid water at -30°C) or any outcome judged improbable is determined to be "non-behavioral". While this can be a subjective process, the originators of this approach point out that there are several instances in which statistical choices are made on partially subjective grounds (e.g., variables included in a model, priors in

Bayesian inference) and that viewing this step in a similar light is warranted because it results in a useful and flexible process (Beven & Binley, 2014).

The classification of behavioral vs. non-behavioral can be based on subjective judgments or using some other more objective values such as Nash-Sutcliffe index. The result is that a proportion of models may be eliminated from the analysis as non-behavioral because they are not viewed as well representing the data. These non-behavioral models are then given a value of zero likelihood of occurring. The second step, after eliminating non-behavioral models, is to weight the remaining behavioral models by their respective likelihood values. These weights are then used to quantify how likely a given output value is and are rescaled so that they can be used to create a weighted cumulative density function (CDF) for a particular output parameter of interest. This method, while very popular, has been criticized as subjective and non-principled because the "likelihood" is not a true likelihood in the sense that it does not adhere to traditional likelihood characteristics but instead functions as an informal method of weighting evidence (Beven, 2012; Wainwright, Finsterle, Jung, Zhou, & Birkholzer, 2014). This is because the errors in this model are thought of as a combination of aleatory (statistical variation) and epistemic (lack of knowledge about the given phenomenon and process). The contention of the creators of the GLUE methodology is that treating epistemic uncertainty as aleatory uncertainty is unwise in that it leads to over confidence (overly "peaked" likelihood functions) in model parameter estimates and poor prediction despite resulting in a "true likelihood" (Beven, 2012; Beven & Binley, 2014). This model framework is significant because it allows for subjective judgements regarding the range of the likelihood function that should be explored. This aspect is related to the approach of FPE analysis in that reasoned but subjective decisions must also be made regarding the area of the fit function surface that should be considered (i.e., magnitude of model fit perturbation). Secondly, this approach has advanced the notion of model Equifinality—that their are several different models that can be used to equally predict an outcome.

While these methods are frequently called global—they only represent a portion of the total theoretical input factor space. For instance, in the process of undertaking a UA and SA, the distribution of each factor must be determined. If an input variable X_4 is thought to be distributed as $N(0, 1)$ then the total factor input space is defined in relation to this assumption. This is an important point because it illustrates the notion that although this approach is "global," the analysis must be limited to a subset of the potential universe of input factor spaces. This maintains the possibility for a useful analysis. If the UA / SA input factors were able to vary across a large (potentially infinite) range than the range of output variables would also be infinitely large and the result of every UA / SA would seem to lead to a hopeless and unhelpful conclusion; namely that any output value is possible. Secondly, acknowledging that the space is limited also highlights that judgments must be made to limit the analysis bounds for practical computational reasons, and not that all input values are equally likely. While referred to as "global" these approaches refer to a subset of the input spaces viewed as reasonable.

Many of the methods reviewed may seem to share commonalities with fungible parameter estimate analysis, however, the two fields in which these methods have been developed seem to operate relatively independently. GLUE, and many other measures are used in the modeling of hydrology data and other complex physical systems (e.g., climate change). Whereas examining fungible parameter estimates comes from a structural equation modeling (and linear regression) framework and those practitioners who originate primarily from a social science background. This distinction in fields is likely partially responsible for the differences in approaches that are possible (or common). In physical based systems, as in hydrology, judgments based on what represents a probable value are easier to justify. For instance, many basic science functions and scientific laws are better known for physical sciences than is currently known in the social sciences. Water does not exist as a liquid at high temperatures (ignoring pressure for simplicity) so it may be reasonable to restrict the theoretical range of lake-water temperature readings to exclude

those temperatures in which water is not a solid. While these judgments are theoretically possible in the social sciences they are more difficult to make, especially when the relation of interest are unobservable (i.e., latent) constructs. In addition, many fields develop their own methodology, although often similar to each other that results from a type of academic isolation. For instance, statistical methods for accommodating clustered data differ between departments despite a similar purpose (McNeish, Stapleton, & Silverman, 2017). The weaker theoretical understanding (i.e., when compared to more mature physical science fields), and academic siloing, have contributed to different approaches to sensitivity analysis between fields. While this may seem to be a negative point against studying latent constructs, the weaker theoretical understanding constitutes an argument for using methods that may be more applicable given the research context. Analysis of fungible parameter estimates originates primarily from a social sciences background that utilizes linear regression, and structural equation modeling frameworks. The types of sensitivity analysis relating to these types of work is now discussed.

2.3 Sensitivity Analysis and Related Techniques in SEM

There are several different aspects that could be considered as part of a sensitivity analysis. It is up to the investigator to determine which factors are reasonable and should be investigated for an uncertainty and sensitivity analysis. In a structural equation modeling framework (SEM) this type of analysis, if done, focuses on case and model selection.

2.3.1 Case Diagnostics

Generalized versions of influence and outlier status have been adapted for an SEM framework. These case diagnostics can be used to determine the effect of particular cases on individual parameter estimates. In this way, researchers may identify data that may have been recorded incorrectly or that might be investigated further. When SEMs are simplifications (i.e., positive degrees of freedom) of the covariance structure they aim to

reproduce, the degree to which they fit the data can often be assessed. In this over-identified modeling context, case diagnostics can also be used to determine whether individual cases are contributing to poor (or improved) model fit rather than each case contributing more or less evenly. This assists in identifying cases in which a poorly fitting model (as identified by global fit measures) may fit well but for a few very atypical cases (Pek & MacCallum, 2011). For instance, those cases that disproportionately contribute to poor model fit may be individuals from a differing population (i.e., older participants in a study of college-aged participants).

2.3.2 Model-Specification

The study of parameter stability and fungible parameter estimates is also thought of as a type of sensitivity analysis. Sensitivity analysis is a general term that refers to determining the impact of potentially any factor that has an effect on the output of a model (e.g., modeling choices, data sub-setting, distributional assumptions etc.) (Saltelli et al., 2008).

More recently, Harring, McNeish, and Hancock (2017) argued that a sensitivity analysis should be undertaken regarding potential external model misspecification (i.e., exclusion of variables from the model). In their approach, phantom variables, variables in which data cannot be obtained or are theoretical, are specified so that one can test the effect of their potential omission on model estimates under a variety of circumstances (e.g., varying parameter strengths etc.). When a model is relatively unchanged by the addition of a phantom variable, we may have more confidence in the results, and conversely we are less confident when the results are appreciably different upon the addition of a phantom variable(s). This is in accordance with a growing understanding that researchers should consider as part of our normal practice an exploration of the multiple factors and the degree in which they affect our results. Similarly, fungible parameter estimates relate the sensitivity of parameter estimates to changes in model fit (i.e., the parameter change needed to result in a specified decrement in model fit).

2.4 Model fit

As discussed elsewhere fungible parameter estimate analysis is in part a sensitivity analysis to model fit. FPE analysis examines what the shape of the likelihood function indicates to researchers who are interested in evaluating the uncertainty in their models. The entire likelihood function is not considered, however, but rather the section that characterizes different solutions under very similar model fit. FPE analysis examines the amount of uncertainty in model parameter estimates by determining how different model solutions might be for a given model. An extremely relevant question is then how is model fit determined? Recognizing that two models can be reasonably considered—essentially equally fitting—is of much importance in accepting the FPE analysis framework. Understanding how model fit is assessed also helps to choose the benchmarks (i.e., the particular index of model fit). Secondly, it may help inform researchers in choosing the degree that the likelihood function should be considered part of the analysis as indexed by the researcher's choice in model fit perturbation.

Most generally, model fit is the degree in which statistical models correspond with a given dataset. As statistical models are simply mathematical equations, model output will always follow model input. This output will often be values that should not be interpreted (i.e., equations do not differentiate meaningful output). To what extent should model output be taken as meaningful? Model fit indices represent one method in which the model output can be evaluated. The purpose of a model is to represent data (i.e., the given system) and a method of evaluating whether this is accomplished is measured by model fit indices. There are several methods of evaluating model fit. It is important to begin with an analysis of how model fit indices are calculated.

Assessing data model fit is an important tool in assessing whether a statistical model is a reasonable representation of the underlying data. In multiple regression analysis, statistical models are often designed as just-identified (i.e., zero degrees of freedom) and therefore necessarily have one perfectly fitting solution (e.g., ML discrepancy function is

zero – see Equation 2.5 for definition). This type of configuration provides no opportunity for assessing model fit. Whereas designing a model so that it is not saturated (i.e., positive degrees of freedom) allows for such an examination. Indices of data-model fit allows for individual models to be falsified, while those with sufficient data-model fit may continue to be considered.

Inducing model constraints necessarily makes the model an approximation, a simplification of the correlations observed in the data, and is unlikely to fit the data perfectly. Designing a model so that it is not saturated (i.e, positive degrees of freedom —an over-identified model) allows for an examination of model fit. Therefore, data model fit is an important tool in assessing whether a statistical model is a reasonable representation of the underlying data.

There are a multitude of model fit indices used within the SEM framework. While different indices can represent very different philosophical approaches to assessing model fit, those within the same class can be quite similar. While FPE analysis is not a statistical "test," at least from a traditional NHST standpoint, it is relevant to review how these different indices benchmark model fit. It is based on these model fit differences that parameter estimates can be characterized as fungible or not.

2.4.1 Model Discrepancy Function

At the most fundamental level, parameter estimates are found after a discrepancy function (e.g., based on the log-likelihood) has been minimized (see Equation 2.5 for a ML discrepancy function used in a SEM context),

$$\hat{F}_{ML} = log|\hat{\boldsymbol{\Sigma}}| + tr(\boldsymbol{S}\hat{\boldsymbol{\Sigma}}^{-1}) - log|\boldsymbol{S}| - p. \tag{2.5}$$

Equation 2.5 represents the model discrepancy function for maximum likelihood estimation in SEM and assumes multivariate normality of the data (Bollen, 1989). Here $\mathbf{S}$ and $\boldsymbol{\Sigma}$ represent the observed and model implied covariance matrices, respectively, and, p

denotes the number of variables. The most elementary unit of model fit may be conceptualized as the fit function value (i.e., discrepancy function value). A value of zero would indicate a perfectly fitting model to the data, thus, the discrepancy function is a measure of absolute model fit. MacCallum, Lee, and Browne (2012) define fungible parameter estimates using two different methods of indexing model fit. First, by using percentages of the raw function values F, and secondly using the root mean squared error of approximation (RMSEA) (Browne & Cudeck, 1992). RMSEA is defined below in Equation 2.7.

2.4.1.1 Chi-Square. The Chi-square test is thought of as the most basic measure of model fit and is utilized as part of many other indices of data-model fit. As is shown in Equation 2.6, the χ^2 value is a function of the maximum likelihood fit function value, F_{ML}, and is defined as

$$\chi^2_{model} \approx T = (n-1)min(F_{ML}), \tag{2.6}$$

where T is the model test statistic that is approximately chi-square distributed given multivariate normality, and n is the number of observations. The Chi-square is a measure of absolute model fit. The null-hypothesis, (H_0), states that the model is perfectly fitting (i.e., zero difference between $\mathbf{S}$ and $\mathbf{\Sigma}$).

To provide context, this approach has been criticized, by some, largely for two reasons: 1) larger sample sizes result in even the most trivial difference between $\mathbf{S}$ and $\mathbf{\Sigma}$ being characterized as poorly fitting (rejecting the null hypothesis) and; 2) because structural equation models are often thought to be simplifications of complicated latent processes it is unlikely that one would expect a perfectly fitting model (Steiger, 2007). This criticism is acknowledged because it may seem a poor decision to utilize an absolute fit function given that in SEM we often do not expect perfectly fitting models.

2.4.1.2 Absolute Model Fit in FPE Analysis. Within an FPE analysis framework, the discrepancy function has been utilized as a method of adjudicating whether

two parameter estimates equally well fit the data. This approach is different from that employed typically by an absolute fit measure. Typically, only the MLE is evaluated, however, in FPE analysis, the purpose is to evaluate the MLE in comparison to other, potentially, similar candidate values. In this respect, utilizing an absolute measure avoids some criticism that it typically might be associated with.

However, previous usage of absolute indices of model fit have had at least two weaknesses that must be overcome. In FPE analysis, the function of fit index is to serve as a method of comparing two candidate values and determining whether they fit the data equally well. Previous research has suggested that parameters resulting from small perturbations of 1% or 5% should be considered as substantively equally fitting (Pek & Wu, 2018). However, there is no substantive interpretation in the unitless function values to base these suggested percent changes. While these suggestions of 1% or 5% seem intuitively small, the suggestion lacks interpretability. Secondly, utilizing percent changes of the discrepancy function was also suggested because it was thought as not being affected by sample size. Because other fit indices are explicitly a function of sample size, Pek and Wu (2018) suggest using a percentage of the discrepancy function. "An alternative perturbation to define FPEs, [fungible parameter estimates], which is explicitly free from sample size and degrees of freedom, can be applied directly to the sample discrepancy function value...". However, the discrepancy function *is* dependent, in part, on sample size, then so too is the percentage based upon it. While the magnitude of the discrepancy function is usually of little interest, its magnitude is dependent on sample size via the value and curvature of the log-likelihood. The increasing sample size results in more information (i.e., a more peaked likelihood function) when compared to smaller sample size. So despite the contention that the discrepancy function is "explicitly free from sample size," it is not. Instead, Ceteris Paribus, as sample size increases the range of FPEs will always be smaller than the range given a smaller sample size. This seemingly contradictory result can be seen in the authors results displaying the FPE range shrinking as sample size increases (Pek &

Wu, 2018, p. 35).

2.4.2 Parsimony Fit Indices

The Root Mean Squared Error of Approximation (RMSEA) (Steiger, 2016; Steiger & Lind, 1980) is a common measure of model fit and has been previously used in order to calculate FPEs. RMSEA penalizes model complexity by defining the best fitting models as ones that best describe the data with few parameters. RMSEA is defined for samples in Equation 2.7,

$$RMSEA = \hat{\epsilon} = \sqrt{max\left\{\frac{\hat{F}_{ML}}{df} - \frac{1}{N-1}, 0\right\}}, \tag{2.7}$$

where $\hat{F}_{ML}$ is defined in Equation 2.5, df denotes the degrees of freedom, N is the sample size, and $\hat{\epsilon}$ represents the value of RMSEA at the MLE. The $max()$ function takes the maximum of its two arguments. Equation 2.7 represents the truncated RMSEA for samples (Browne & Cudeck, 1992; Steiger, 2000). Unlike values of the discrepancy function, RMSEA values are typically utilized to adjudicate model fit. As a guideline, RMSEA values of .06 are interpreted as representing satisfactory model fit (Hu & Bentler, 1999). The judgment to consider values under .06 is supported by simulation study, but similar to many statistical practices, is based to some extent on judgment. There is considerable debate regarding what constitutes "good" fitting models (Steiger, 2007; Goffin, 2007). These discussions are vital and the presence of these disagreements does not negate the use of RMSEA or similar indices for use in calculating FPEs. Instead, this information should be used to best adjust the size of the decrement in model fit for a given scenario (i.e., given simulation research—what constitutes good fit for this scenario?). This type of research helps to better inform the interpretation of what constitutes "good" model fit for a given scenario and strengthens researchers interpretation of FPE analysis under a variety of scenarios.

In addition, RMSEA has the advantage of being less influenced by sample size than

does the discrepancy function values. This is useful because FPE analyses is intended to quantify uncertainty as indexed by model fit and not necessarily from sampling uncertainty. Also of utility, this relative unreactiveness to sample size allows RMSEA to serve to differentiate between the two aforementioned forms of uncertainty.

2.4.3 Incremental Fit Indices

Incremental fit indices are methods of assessing model fit between two candidate models. From this framework, a null model is compared against a candidate test model. At first glance, FPE analysis may seem conducive to this type of model comparison by reasoning that FPE analysis is a method of comparing multiple model estimates. However, as noted by T. Lee et al. (2017), incremental indices are unlikely to be appropriate in FPE analysis precisely due to its comparison with a null model. In this modeling framework, a null model must be nested within all potential models of interest tested (Widaman & Thompson, 2003). If the null-model used does not meet this requirement, a global comparison of potential models is not justified. A model is nested within another if it can be found by imposing a constraint on the original model. Typically, null models are conceptualized, as one in which all variables are assumed to be uncorrelated, however, this is not necessarily the case in particular modeling frameworks (e.g., latent growth modeling).

2.4.4 Akaike Information Criterion (AIC) and Related Measures

The Akaike information criterion (AIC) is an information based measure used to evaluate the suitability of set of candidate models (Akaike, 1974). This type of method differs from absolute and incremental indices used in SEM and may be useful when used in a FPE analysis framework for a few reasons.

AIC is a index of the relative model fit used in model selection. AIC represents an extremely popular index that is used in many fields. According to Google Scholar as of 2018, the original journal article "A New Look at the Statistical Model Identification" (Akaike, 1974) by statistician Hirotugu Akaike has been cited over 42 thousand times and represents one of top 100 cited papers. This method, however, is not typically one of the

most popular indices reported or used in a SEM framework. One reason for this may be that AIC focuses on model selection and comparison whereas SEM is often conducted under a confirmatory framework in which a particular model can either be falsified or not. AIC cannot be used in order to verify that a particular model is well fitting, but only in comparison to other candidate models. However, the process of undertaking an analysis of FPEs can be conceptualized as a comparison of models. This is because different candidate θ_i can be thought of as different candidate models representing the data. Typically, θ_{ML} is taken by researchers as the best of all candidate model parameters. How does the set of maximum likelihood parameter estimates however compare to other models (θ_i). To understand the potential utility of this index within a FPE analysis framework, a short review is provided next.

AIC is a measure of relative model fit between models. It is based on information theory and its underlying theory constitutes an extension of likelihood-theory (Akaike, 1998). In particular, AIC is an estimate of Kullback-Leibler (K-L) distance (Akaike, 1974; Kullback & Leibler, 1951). K-L distance is a measure of discrepancy between two probability distributions. It was formulated as a method of quantifying the amount of information lost given a probability function π_i versus the actual probability distribution function p_i. Kullback and Leibler originally intended to quantify the amount of information loss (Shannon, 1948) by the inclusion of extraneous information. Yet, the notion of K-L divergence applied not only in communication theory (i.e., computer science) but also as a more general method of comparing how similar two distributions are (and not necessarily only regarding the minimum number of bytes to convey a message).

Using K-L distance as a measure to determine how far a specific distribution (i.e., our model) is from full reality (i.e., the true model) might be useful, however, it had a major disadvantage of requiring both knowledge of the true distribution (the referent distribution p_i) and the test distribution (π_i). Akaike's contribution was that K-L distance could be estimated using the referent estimates at the maximum likelihood estimates. Specifically,

the parameter estimates found by maximizing the log-likelihood also minimized the K-L divergence between the referent and full-reality model.

2.4.4.1 Reviewing AIC Interpretation. AIC is a measure of divergence between full reality and a particular model. The values of AIC are not useful and instead, AIC values are intended to be compared between multiple models. Models with the lowest value are the best representative of the full generating model. While singular AIC values do not have an interpretation there are guidelines to interpreting multiple models. Burnham, Anderson, and Burnham (2002, p. 70) suggest calculating $\Delta_i = AIC_i - AIC_{min}$. Where the ith model represents different candidate model AIC values and AIC_{min} is the lowest of all AIC model values. AIC value increases from the minimum between 0-2, 4-7, and greater than 10, have interpretations of "Substantial", "Considerably less" and "Essentially None" levels of support when compared to the best representative model (i.e., AIC_{min}).

AIC is an estimate for K-L divergence. Specifically, the maximum likelihood estimates were shown to minimize the K-L distance. Any other θ value(s) representing the model will then necessarily represent non-minimized K-L distances. Typically, this non-minimized distance would not be of interest but in this case we are interested in understanding to what extent alternative models can be differentiated in terms of their level of support as measured by K-L distance, which is defined as

$$I(f,g) = \sum_{i=1}^{k} p_i * log(\frac{p_i}{\pi_i}) \tag{2.8}$$

$$AIC = -2log(\mathcal{L}(\hat{\theta}|y)) + 2K \tag{2.9}$$

Equation 2.9 represents Akaike's Information Criterion (AIC). The first term represents the value of the log-likelihood, typically measured at the maximum (where the K-L distance is minimized). Here the second term, k, equals the number of estimated parameters. Lower AIC values are preferred thus the second term can be thought of as a

penalty for model complexity.

A modified version of AIC was also formulated to better accommodate small sample sizes. The formula for AIC_c is shown in Equation 2.10

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}. \tag{2.10}$$

AIC_c adds an additional term that adjusts the value of the index for small sample sizes. However, despite this adjustment AIC_c should be interpreted in the same manner as AIC. It is the recommended index in cases in which the sample size is not sufficiently large relative to the number of parameters estimated (e.g., ratio of n/K less than 40, Burnham et al., 2002). This adjustment would then seem especially pertinent given an SEM framework in which sample sizes are frequently small, and the estimated number of parameters relatively large. While typically true, for the purposes of an FPE analysis, the non-adjusted AIC is preferred. This is because FPE analysis considers different models consisting of a multitude of θ, values but holds constant the number of parameters and sample size. Thus, the AIC_c adjustment is the same for both the decremented (i.e., FPE models) and the MLE model. The same relative difference is maintained and no adjustment is needed. The relative difference is the relevant quantity for a valid interpretation of AIC.

2.4.4.2 Bayesian Information Criteria. Bayesian Information Criteria (BIC) is another information criteria popular for model selection. Similar to AIC, BIC has had relatively little usage in an SEM framework compared to non-information criteria based measures. BIC, however differs from AIC in that it is not a measure of K-L distance. Instead, BIC is a method for model selection based on the notion of selecting the model with the highest posterior probability given the data (Schwarz, 1978). This method is effective at selecting the optimal model given that the "true" model is one of the candidate models under consideration (Burnham et al., 2002). BIC has also been primarily justified using asymptotic theory and it may not perform well in conditions other than those with very large sample sizes (Burnham et al., 2002; Huang, 2017; Schwarz, 1978). These two

conditions (i.e., large sample size, true model inclusion) make it difficult to prefer BIC in an SEM framework.

BIC is formulated as:

$$BIC = -2log(\mathcal{L}(\hat{\theta}|y)) + log(n) * K.$$

Here it can be seen that despite their philosophical differences, AIC and BIC differ only by the second term. As noted by Schwarz (1978), this term has the effect of preferring models that have fewer parameters when compared to AIC. It may then seem that BIC is preferable in circumstances in which one aims to adopt a more conservative approach. That is, BIC has a stronger "penalty" for overparameterization than does AIC. However, it is of interest that the $2K$ parameter in AIC was not added as an arbitrary preference for or against more parsimonious models. Rather Akaike realized the MLEs were a biased estimate of K-L distance approximately equal to $2K$ (Akaike, 1974; Burnham et al., 2002). While both AIC and BIC are popular and principled methods for model selection, AIC is the preferred method for the aforementioned reasons.

Data-model fit allows for individual models to be falsified, while those with sufficient data-model fit may continue to be considered. Statistical approximations of complicated phenomenon are difficult to prove correct, yet the possibility for falsification strengthens one's conclusions. The falsifiability of a theory (i.e., model) is a cornerstone of the scientific method. In addition, sufficient data model-fit is usually considered evidence for favoring one statistical model over another. However, the practice of selecting between a very limited number of models based solely on model fit is challenged by the knowledge of a large number of equivalent models. Overall model fit provides an incomplete picture, and therefore additional information is necessary when making decisions regarding model evaluation.

2.5 Model Indeterminacy in SEM

2.5.1 Equivalent Models

Rather than providing evidence for a particular theory, model fit should instead be thought of as a tool to refute hypotheses when no other model is being considered. Work by MacCallum, Wegener, Uchino, and Fabrigar (1993) indicates that the number of rival hypotheses that support any given data set can often be very large. Originally, model comparison was primarily limited to nested models that could be compared with a likelihood ratio test; later, this work was extended to include models that were not nested, and included a mean-structure or multiple groups (Levy & Hancock, 2011). Furthermore, work done by Lai, Green, and Levy (2017) argues that in addition to questioning whether models should be considered as equivalent, or nested, before data is collected, researchers should also take into account how potential covariance patterns can result in effectively equivalent models. There are multiple benefits of using this approach. First, the approach makes salient the potentially large number of equally or similarly fitting models that often have different or opposite interpretations given data. This quality is important because it helps the researcher recognize the degree to which they must rely on theoretical rational to justify a particular interpretation.

Second, in some cases, experiments can be designed in order to falsify rival hypotheses, a task (Lai et al., 2017) argue is made easier once the full landscape of possible alternative explanations is explored. An understanding that a potentially large numbers of equivalent models may also equally describe the data demonstrates that while model-fit is part of model evaluation, it should not be used as the sole criterion for evaluation.

2.5.2 Causal Search

Reminiscent of the expansive approach used by Lai et al. (2017) is the method taken by those who employ what are called causal search algorithms (Lagani, Triantafillou, Ball, Tegner, & Tsamardinos, 2016; Spirtes, 2010). Causal search algorithms work in a manner opposite to that which researchers are accustomed in many social science disciplines.

Typically, theory drives the model, the model is estimated given the data, then the model is either falsified or not (i.e., model fit). Some at this point– controversially– make model changes to improve model fit (e.g., Lagrange Multiplier test, Wald Test). Making any changes to improve model fit, however, is inadvisable due to the risk of overfitting the data and/or improperly interpreting the results (i.e., results may not be replicable because they are sample specific). While post-hoc changes to improve model fit may seem similar, and perhaps even more conservative, to an exploration of equivalent models, or causal search, they are philosophically different. An exploration of equivalent model space and the approach taken by causal search help us understand the degree to which the data can confirm a given hypothesis. Compared to post-hoc changes that can, unfortunately, amount to a futile exercise in overfitting rather than a confirmation of a causal hypotheses. Causal search algorithms work only after data have been collected and by implementing rules for *d*-separation (Pearl, 2009). Applying these rules for conditional independence, causal search algorithms attempt to determine the full set of causal graphs that might have generated a given data set. Using this approach, common features between the causal graphs can indicate potential targets for experimentation or assist in removing large sets of hypotheses that *are not possible* given the data. This approach is similar to that of the equivalent model search discussed above, or to an analysis of fungible parameters, in that it provides information regarding the scope of potential models (i.e., causal hypothesis), and parameter estimates that may be possible, respectively. In either case, the investigator has more information that is available to help evaluate the plausibility of a particular or set of hypotheses. As described below, an examination of fungible parameter estimates is another approach for gathering information regarding model-fit, that alone cannot prove the validity of a model. However, fungible estimates can be used to support or weaken the subsequent interpretation of parameter estimates by demonstrating the degree in which the model represents the data.

Overall, both model-fit and sensitivity analysis are two important methods in which

models are evaluated. By examining model fit of a single model (or as informed by methods developed by the equivalent models literature or causal search algorithms) researchers gain a better understanding of whether a model represents the data or if other potential models should be considered. For example, a researcher may ask, can both models equally describe the data? By conducting sensitivity analyses researchers gain more information regarding uncertainty contained in their results. In line with both of these approaches, an analysis of fungible parameter estimates is a type of sensitivity analysis regarding the changes in parameter estimates and model-fit.

2.6 Fungible Parameter Estimates

2.6.1 Observed Variables

Fungible estimates have been studied primarily in the context of linear regression analysis (Hoerl & Kennard, 1970; Koopman, 1988). Under a linear regression framework, research has suggested that small decrements in R^2 may be associated with equal or even improved prediction performance when compared to the least squares solution under predictor multicollinearity (e.g., Wainer, 1976, 1978). More recently, methods for identifying fungible parameter weights in the context of multiple regression (Waller, 2008; Waller & Jones, 2009) and logistic regression (Jones & Waller, 2016) have been proposed as a method of studying parameter estimate stability with the primary difference being the estimation method used. OLS was utilized in the context of linear multiple regression while maximum likelihood was employed for estimating logistic regression models. Each modeling context used the same steps in deriving sets of fungible coefficient estimates. These steps are outlined below.

Waller's method (Waller, 2008; Waller & Jones, 2009) for calculating fungible regression weights for a linear regression model follows three primary steps: 1) identify a measure of decrement of model-data fit, 2) translate that measure to potential fungible regression weights, and 3) compute sets of fungible regression weights. A standard linear regression model can be used to illustrate the main features of the method. For outcome

variable, Y, let, $\mathbf{y}$ be a vector of instantitations of Y following

$$\boldsymbol{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{2.11}$$

where $\mathbf{X}$ is an $n \times (p+1)$ matrix of independent variables, $\mathbf{b}$ is a $(p+1) \times 1$ vector of regression coefficients and $\mathbf{e}$ is an $(n \times 1)$ vector of errors. The method of ordinary least squares minimizes the residual sum of squares. Any other composite set of weights, say $\mathbf{a}$, will necessarily explain less variance as measured by the coefficient of determination, R_a^2. Model fit cannot be assessed in a just-identified model, however, R^2 is often used as a fit measure in this context. Waller (2008) demonstrated that for any decrement of R_b^2 from the least squares solution resulted in an infinite number of interchangeable regression weights (or coefficients). Waller defined the relation between a fungible set of coefficients and decrement of R^2 as

$$r_{\hat{y}_{\boldsymbol{b}}\hat{y}_{\boldsymbol{a}}} = \frac{\boldsymbol{a}'\Sigma_{\boldsymbol{X}}\boldsymbol{a}}{(\boldsymbol{a}'\Sigma_{\boldsymbol{X}}\boldsymbol{a})^{1/2}(\boldsymbol{b}'\Sigma_{\boldsymbol{X}}\boldsymbol{b})^{1/2}} = (1 - \frac{R_b^2 - R_a^2}{R_b^2})^{1/2}. \tag{2.12}$$

Equation 2.12 defines the correlation between $\hat{y}_{\boldsymbol{b}}$ and $\hat{y}_{\boldsymbol{a}}$ where $\hat{y}_{\boldsymbol{b}}$ represents the fitted values using the weights calculated using the least squares criterion and $\hat{y}_{\boldsymbol{a}}$ represents the fitted values using a set of sub-optimal regression weights. Importantly, the relation between the sub-optimal regression weights, $\mathbf{a}$, and optimal (i.e., as judged by the Least Squares criterion) regression weights, $\mathbf{b}$, is also equal to a particular decrement in R_b^2 computed using the OLS solution.

From here, Waller derived an expression to solve for an infinite set of fungible parameters using the spectral decomposition of the covariance matrix of $\mathbf{X}$ as

$$\mathbf{a}_i = (\boldsymbol{\sigma}'_{\mathbf{X}_y}\mathbf{V}\boldsymbol{\Lambda}^{-1/2}\mathbf{k}_i)\mathbf{V}\boldsymbol{\Lambda}^{-1/2}\mathbf{k}_i, \tag{2.13}$$

where $\mathbf{V}$ is a $(p \times p)$ orthogonal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is a $(p \times p)$ ordered diagonal matrix of eigenvalues from the spectral decomposition, $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$, and where $\boldsymbol{k}_i = r_{\hat{y}_{\boldsymbol{a}}\hat{y}_{\boldsymbol{b}}}\boldsymbol{u} + \sqrt{1 - r^2_{\hat{y}_{\boldsymbol{a}}\hat{y}_{\boldsymbol{b}}}}\mathbf{U}\mathbf{z}_i$. Once $\boldsymbol{k}_i$ unit length vectors had been identified, Waller (2008) showed that i sets of fungible parameters, $\boldsymbol{a}$, could be generated at a prespecified

value of R_a^2 (see Waller, 2008, for a more comprehensive technical discussion). Waller does not, however, provide specific guidance regarding how to interpret the fungible parameter estimates. Fungible parameter estimates are in the same scale as the original estimates so it might be assumed that researchers should, based on their substantive knowledge, determine whether the range of fungible estimates is unacceptable.

2.6.2 SEM Generalization

The methodological procedure for calculating FPEs in an observed regression framework has since been generalized to an SEM framework (MacCallum et al., 2012). SEM is framework commonly used in several social sciences (e.g., Psychology, Education, Economics) and provides a flexible framework capable of modeling latent (i.e., unobserved) variables.

Once a model has been specified, estimation approaches such as ML are often utilized to obtain a set of estimates of unknown parameters (e.g., regression slopes, latent variable path coefficients, residual variance components) for a given model. In latent variable software programs like M*plus* or the `laavan` package in R, this estimation is accomplished through iterative procedures that minimize the fit function, F_{ML}, where $\hat{\boldsymbol{\theta}}$ represents the set of ML estimates that are obtained upon convergence to the global minimum of the fit function, $\hat{F}_{ML}$. This iterative process, and the function that produces it, provides researchers a wealth of information that can be used for model evaluation, despite this fact all but the final iteration are typically unknown to most researchers. Through an examination of the range of these approximately equivalent parameter estimates (i.e., fungible parameter estimates), researchers can obtain information regarding the degree in which each parameter describes the data and overall model plausibility.

In an SEM framework, MacCallum et al. (2012) suggested a method for identifying fungible parameters whose main points are now summarized. As mentioned previously, $\hat{\boldsymbol{\theta}}$ represents the value of the parameters at the optimized value of fit function $\hat{F}$. A perturbation of any parameter in $\hat{\boldsymbol{\theta}}$ would result in an a slightly reduced level of fit.

MacCallum et al. (2012); T. Lee et al. (2017) then demonstrated that fungible parameter estimates can be identified in three steps. First, ML estimates of parameter vector $\hat{\boldsymbol{\theta}}$ are obtained by fitting a model to data (Equation 2.5). Second, a particular amount of parameter perturbation is chosen. For this important step, MacCallum et al. (2012) defines fungible parameter estimates using two different methods of indexing model fit. First, by using decremented values of the raw fit function F directly (e.g., 10% increase in the value of F), and secondly using the root mean squared error of approximation (Browne & Cudeck, 1992). For RMSEA, model decrement is defined as $\hat{\epsilon}^* = \hat{\epsilon} + \tilde{\epsilon}$. Where $\tilde{\epsilon}$ is the size of the decrement in model fit and $\hat{\epsilon}^*$ is the new perturbed value of RMSEA. Estimates that have an RMSEA equal to that defined by this definition are considered fungible (in accordance with Equation 1.1). This definition along with Equation 2.7 is then used to solve for the adjusted target fit function value, $\hat{F}^*$, that defines the fit at which estimates are considered fungible. The following two equations summarize the approach for identifying FPEs.

$$\hat{\boldsymbol{\theta}}_n^* = \hat{\boldsymbol{\theta}} + \boldsymbol{\kappa_n} \boldsymbol{d}_n \tag{2.14}$$

$$F(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}} + \boldsymbol{\kappa}_n \boldsymbol{d}_n), \mathbf{S}) - \hat{\boldsymbol{F}}^* = 0. \tag{2.15}$$

As shown in Equation (2.14), the new set of fungible parameter estimates, $\hat{\boldsymbol{\theta}}^*$, can be found by solving for the scaling constant $\boldsymbol{\kappa}_n$ given the magnitude of the perturbation selected based on the value of RMSEA. The authors solved for $\boldsymbol{\kappa}_n$ in Equation (2.15) using a root solving algorithm. Here, $\boldsymbol{d}_n$ represents a unit length vector. The process is repeated to calculate any number of fungible parameter estimates for a given parameter. This approach has several limitations First, it involves searching for fungible parameters for only a limited set of "focal parameters" while fixing the remaining parameters at their ML estimates. As the number of focal parameters increases, so too do the number of dimensions that vectors $\boldsymbol{d}_n$ must search, greatly increasing the computational burden. The authors discuss only examples with two or three focal parameters partly due to this

limitation. A second important limitation of this method is that because the majority of parameters are held constant, it is not possible to explore the full range of the fungible parameter space. That is, if all parameters were able to vary simultaneously the fungible parameter space would likely be larger for focal parameters (see T. Lee et al., 2017, p. 11).

In order to overcome this limitation, (Pek, 2012) utilized a modified version of the root solving algorithm employed by MacCallum et al. (2012); T. Lee et al. (2017). This algorithm adds an additional step of maximizing the likelihood of the nuisance parameters at each candidate value of $\hat{\boldsymbol{\theta}} + \boldsymbol{\kappa_n d_n}$. That is, $\hat{\boldsymbol{\theta}}_n^*$ is estimated after holding constant $\hat{\boldsymbol{\theta}} + \boldsymbol{\kappa_n d_n}$ rather than jointly as in the method used by T. Lee et al. (2017). Pek and Wu (2018) further suggested utilizing an alternative algorithm by Wu and Neale (2012). This algorithm is capable of calculating unbiased profile likelihoods (i.e., iteratively re-estimated nuisance parameters) in a more efficient manner. Despite this computational improvement this method is only appropriate for calculating a limited number of focal parameters.

2.6.2.1 Multi-Dimensional Approaches.

Research has so far focused on cases in which there are relatively few dimensions searched for fungible parameter estimates. S.-Y. Lee and Wang (1996) demonstrated an approach to identifying sensitive parameters in structural equation models. Rather than perturbing RMSEA (or another fit measure) and calculating the change in parameter estimates, their approach perturbs the parameter values (i.e., MLEs) and identifies the change in fit. Additionally, these same authors showed that it was possible to use the Hessian matrix estimated at the maximum likelihood solution to identify the eigenvector that results in the largest decrement in model fit. Using this approach, the parameters that resulted in the largest (or least) influence on model fit could be identified. T. Lee and MacCallum (2015) used a similar procedure to show that the process can be extended to identify the degree in which individual parameters are most influential and the degree in which they would need to change in order to meet a pre-specified model fit decrement.

2.6.3 Differences With Other Measures of Uncertainty

Because fungible estimates are typically presented as a range of values representing a type of uncertainty it is important to differentiate them from the uncertainty measured by standard errors. There are a few reasons why examinations of the likelihood function (through the examination of the fungible parameters) are useful and distinct from the information given in standard errors. Confidence Intervals (CIs) provide information about the precision of estimates given a frequentist interpretation. That is, what is the expected proportion of confidence intervals that would contain the true parameter value given repeated sampling (e.g., on average 95 estimates would be within a particular confidence interval given 100 experiments). Whereas, fungible parameter estimates inform researchers to the level of confidence, and more specifically, the amount of support for the MLE given a particular statistical model and data. Importantly, an increasing sample size will always decrease the size of standard errors but this narrowing relationship does not hold for the range of fungible estimates (Pek & Wu, 2018). Because a correctly specified model is a pre-requsite for correct interpretation of MLE and their accompanying standard errors, this provides justification for examining the fungible parameter space before considering standard errors. Secondly, because information (i.e., peakedness of the likelihood function) informs how important a particular parameter is in terms of affecting the model fit, it may also serve as important method of assessing model misspecification (Pek & Wu, 2018). For instance, robust parameter estimates (high information) for structural paths of a latent variable model might indicate that this section of the model fits well whereas, a structural path with a wide range of fungible parameter estimates may indicate a subset of the model is not accurately describing the data and may need to be changed. This, of course, does not relieve an investigator of the duty of pairing statistical knowledge with theoretical knowledge. However, using the likelihood function and fungible parameter estimates as a measure of local model fit may be useful when comparing two different competing models (in conjunction with theoretical considerations), or when conducting an exploratory

analysis for use with further data collection. Finally, because standard errors are calculated using expected or observed Fisher Information (i.e., the Hessian matrix at the MLE) they will not necessarily represent the uncertainty of the entirety of the likelihood function when the function is complex, even under large sample sizes.

Also from within an SEM framework, the distinction between fungible parameters and standard errors has been made explicit by means of demonstrating how they differ from confidence sets (i.e., a multivariate generalization of confidence intervals) (Pek & Wu, 2018). Specifically, the authors demonstrate an analytical relationship and conceptual difference between confidence sets (i.e., confidence intervals for single parameters or confidence regions for multiple parameters) and fungible parameter estimates. This article conceptualizes the distance to the boundaries in profile likelihoods as a type of perturbation (i.e., distance from the MLE) that can be calculated in the same scale as the perturbation used in calculating the distance of fungible parameter estimates. Despite this, the authors point out our interpretation should differ between the two concepts. In addition, the authors also conducted a limited simulation study to demonstrate how confidence sets, and fungible estimates might react to changes in sample size, model fit, and the magnitude of correlations between measured variables. Results indicated that models with smaller magnitude structural paths, and smaller unique variance of measured variables resulted in narrower bands of fungible parameter estimates.

2.6.4 Theoretical Justification for FPE Analysis

Calculating fungible parameter estimates is a method for examining how sensitive parameter estimates are to small changes in model-fit. The method of analyzing these fungible estimates represents a technique that is relatively uncommon for many researchers and therefore justifies the comparison to related techniques to improve understanding. While those who utilize sensitivity analysis or analyze model fit will likely see the similarities with an analysis of fungible parameter estimates. Both a conceptual overview, and technical discussion will assist to further justify the use of FPEs while also

distinguishing the information provided from other techniques.

The likelihood reverses the conditional relationship described in probability by treating data as a constant and θ as the random variable. The likelihood of a set of parameters θ is conditional on observed data, x. Likelihoods are not probability statements about θ but can be used as part of Bayes Theorem to calculate probabilities regarding θ (Gelman et al., 2013). A likelihood function can be defined as follows

$$L(\theta) \equiv L(\theta|x) = c(x)Pr(x|\theta) \propto Pr(x|\theta). \tag{2.16}$$

The likelihood function allows for the comparison of the relative likelihood of different candidate values of θ conditional on a given sample of observed data x and is equal to the probability of x, given θ is multiplied by a constant $c(x)$ (Azzalini, 1996; King, 1998, p. 22). Because the value of the likelihood depends on the scale of the data and on the magnitude of the parameters, the absolute likelihood value is not important or interpretable. Instead, only the relative values of a likelihood function should be examined (likelihood function values, unlike a probability density function, do not inherently integrate to one).

The likelihood principle is similar to a Bayesian statistical framework in that it links prior knowledge (i.e., the chosen likelihood model) with collected data (Azzalini, 1996, p. 28). Rather than making a probabilistic claim the likelihood axiom states that all information regarding the relative merits of differing candidate values of θ—that can be inferred by a particular dataset—are contained within the likelihood function, given a particular statistical model (Edwards, 1972, p. 31; Fisher, 1922, p. 326). Edwards (1972) argued that while the likelihood principle lacks a probability interpretation it is nevertheless important because it allows for comparison of the relative support for one statistical hypothesis over another (e.g., $\theta = 3$ vs. $\theta = 3.1$). Similarly, examining the likelihood function through the calculation of fungible parameters is a method of communicating the uncertainty of maximum likelihood estimates without appealing to probability or long run frequencies.

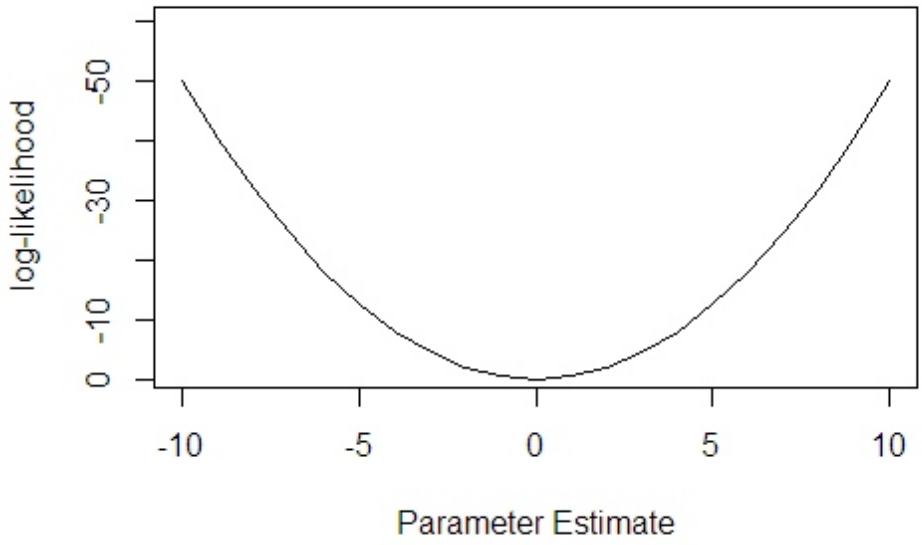

Figure 2.2. Peaked likelihood function: log-likelihood vs. parameter estimates. This example shows when parameter estimates change the model fit (i.e., the log likelihood value) decreases relatively quickly. Here the model fit is sensitive to changes in the parameter estimate.

The shape of the likelihood function contains meaningful information that is gathered by calculating fungible parameter estimates. This point is illustrated by two simple example likelihood functions. The first function, shown in Figure 2.2, is peaked and demonstrates the degree in which model parameter estimates change quickly relative to model fit (i.e., model fit is sensitive to changes in the parameter estimate). Conversely, Figure 2.3 illustrates that different parameter estimates may have a very similar model fit value (i.e., model fit is *not* sensitive to changes in the parameter estimate). These two likelihoods feature the same MLE (i.e., global minimum) but are otherwise different functions. The goal of a fungible parameter estimate analysis is therefore to explore the shape of the likelihood function. Specifically, this is done by calculating parameter estimate values that are very close to the optimal data-model fit (i.e., MLE). Traditionally, the section of the likelihood function explored in FPE analysis represents parameter estimates at only one (non-optimal) value of the model fit. Parameter estimates calculated at this

"slice" of the likelihood function, therefore, equally describe the data and are completely interchangeable (i.e., fungible). It is important to note that while the example functions in Figures 2.2 and 2.3 are useful for descriptive purposes they represent simple likelihood functions and that realistic examples have likelihood surfaces with potentially high order dimensions that are not easily visualized. In addition, determining how to calculate fungible parameter estimates along more complicated likelihood surfaces forms an important area of fungible parameter estimation research.

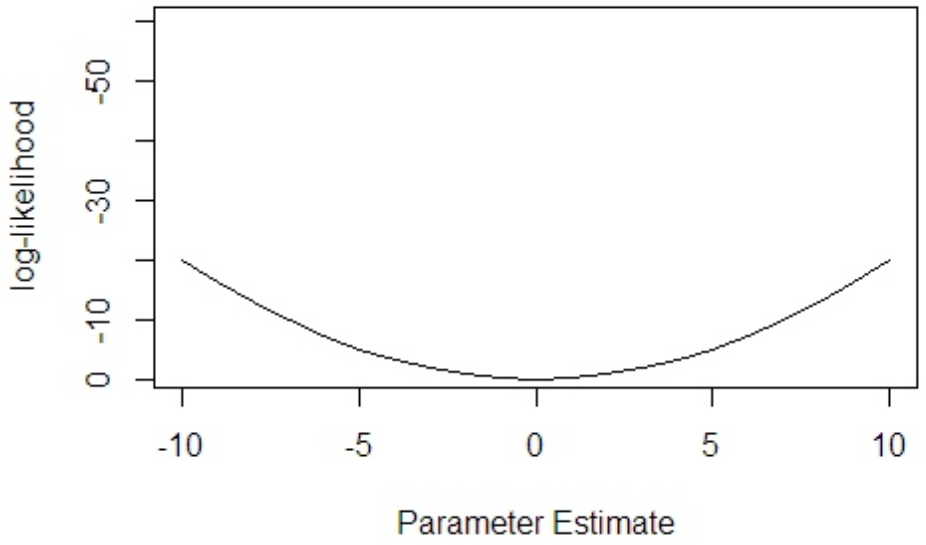

Figure 2.3. Non-peaked likelihood function: log-likelihood values vs. parameter estimates. This figure shows that the parameter estimates can change to a large degree with only relatively small changes in model fit (i.e., the log likelihood value). Here the model fit is relatively insensitive to changes in the parameter estimate.

The shape of the likelihood provides information regarding the level of support for a given hypothesis, θ, compared to other potential values. The importance of the shape of the likelihood function is in accordance with the law of likelihood shown below:

$$\Lambda = \frac{\mathcal{L}(\theta_1|X=x)}{\mathcal{L}(\theta_2|X=x)} = \frac{p_{\theta_1}(x)P(\theta_1)}{p_{\theta_2}(x)P(\theta_2)},$$

where Λ represents the likelihood ratio and θ_1 and θ_2 represent competing hypothesis (i.e., potential parameter values). If Λ is greater than one then there is support for θ_1 relative to θ_2 and, importantly, the magnitude of Λ measures the strength of the evidential support for one θ_1 relative to θ_2 (Hacking, 1965). Part of the justification for utilizing the MLE (over any other value) can also serve as a rationale for interpreting the relative support (i.e., the ratio of values—the shape of the function) of the MLE to other similarly adequate estimate (i.e., fungible parameter estimates). Here $p_{\theta_1}(x)$ and $p_{\theta_2}(x)$ may refer to either probabilities or probability densities. In summary, the relative peakedness (i.e., shape) of the likelihood function will necessarily change the likelihood ratio and indicates that a more peaked shape constitutes stronger support for the MLE compared to other potential estimates. Whereas, a relatively flat surface (in the simple example) constitutes weak support for the MLE compared to other potential estimates.

2.6.5 Generalized FPE Analysis and the PSINDEX

A new generalized method for calculating FPEs by simulated annealing is now introduced. This method remedies some of the previous methods for calculating FPEs. As mentioned earlier, the method for calculating FPEs by T. Lee et al. (2017) (and all others discussed) have several limitations:

1) Each focuses on calculating FPEs for a limited number of "focal parameters". This is problematic because the manner in which uncertainty is manifested in the set of FPEs is not known. That is, while it is thought that FPEs will accurately localize uncertainty to parameters it is not known whether uncertainty will be distributed across multiple parameters or not. The previous approaches obscure our ability to investigate this assumption. It is important not to unnecessarily constrict the results of this analysis to a few parameters when the behavior and interpretation of these results still remains largely unknown.

2) These methods could not overcome this limitation due to being computationally very burdensome. The new generalized global FPE approach calculates unbiased FPE

estimates simultaneously for all parameter estimates (e.g., even with a very large number of estimated parameters).

3) Most importantly, each conceptualizes what an FPE analysis is in an overly restrictive way (i.e., as 'slices' rather than as a general uncertainty and sensitivity analysis technique). It is important to conceptualize FPE analysis as a type of uncertainty and sensitivity analysis technique. Thus, the goal is an exploration of uncertainty in the input factor space and its affect on the output (i.e., model fit) (Saltelli et al., 2008). It is the surface of this input factor space in which an FPE analysis aims to explore. Slices of this surface are approximations of the larger surface (see discussion accompanying Figures 2.6 and 2.7) and are computationally difficult to generate.

Importantly, this method is available for easy use in the psindex package. The method by (T. Lee et al., 2017) is not currently implemented in software. Pek and Wu (2018) method can be calculated using the OpenMx package. This is accomplished by repurposing the R package OpenMX's ability to calculate likelihood confidence regions. However, in order to calculate FPEs one must understand how to properly convert the decrement in model fit to the appropriate "confidence level" or reparameterize the model in order to obtain the correct set of estimates (Pek & Wu, 2015, 2018).

2.6.5.1 Parameter Stability Index. This new method is implemented in the **P**arameter **S**tability **Index** (`psindex`[1]) a new R package that can be used to easily calculate FPEs for interpretation by methodologists and applied researchers (Prendez & Harring, 2019). The `psindex` builds upon the work of a multitude of previous open source software including the popular R package `lavaan` (Rosseel, 2012), and maintains its simple model-syntax conventions. Using `psindex`, researchers can generate fungible parameter estimates for a large subset of models that can be accommodated by the `lavaan` package. Results from all forthcoming examples were calculated using `psindex`.

[1]The current instructions for downloading and using `psindex` can be obtained by visiting `https://github.com/nietsnel/psindex`

2.6.5.2 Simulated Annealing Method. The `psindex` package calculates FPEs using a simulated annealing algorithm. This method is well-suited for exploring the likelihood function and can be used to conveniently calculate FPEs utilizing a method that differs from previous approaches, as discussed in the subsequent sections.

In general, optimization algorithms can be characterized as either deterministic or stochastic. By default, `lavaan`, employs a hill climbing optimization method (i.e., the nlminb function in R) that employs a deterministic search in route to a solution. This type of deterministic optimization algorithm (others include the Newton-Raphson algorithm) is utilized by most SEM programs, with the exception of Bayesian SEM. In contrast, simulated annealing is a stochastic optimization method that works similarly to traditional hill climbing algorithms but can probabilistically accept non-optimal solutions. The probability that a non-optimal solution will be accepted is based on a temperature parameter that starts at a high value and slowly decreases with successive algorithm iterations. While the temperature parameter is high, inferior solutions have a higher probability of being accepted whereas inferior solutions have a lower probability of acceptance as the temperature parameter decreases (Cortez, 2014). Although hill-climbing algorithms can be efficient, they can become trapped as they often do not have a mechanism for climbing out of local optima. Thus, they are also not well-suited for optimizing very complicated functions. This same ability to accept successive non-optimal solutions also gives simulated annealing a scheme for sampling the parameter surface close to the ML estimate that deterministic approaches do not possess.

Below, the process for calculating FPEs for a model when using the `psindex` is enumerated:

1. The end user specifies a structural equation model and the maximum decrement in model fit (e.g., RMSEA $\leq$ .01) to be explored by calculating FPEs.
2. The `psindex` calculates parameter estimates using maximum likelihood and `lavaan`'s default optimizing algorithm (i.e., nlminb function in R). The ML estimate is then used

as the starting values for the simulated annealing algorithm.

3. The `psindex` explores the log-likelihood function parameter space using the simulated annealing optimization algorithm found in the `GenSA` R package (Xiang, Gubian, Suomela, & Hoeng, 2013). The simulated annealing algorithm was created to explore complicated functions in search for a global optima. This capacity is re-purposed in order to provide a method to sample function values that are near the global optimum (i.e., FPEs).
4. Parameter values are stored for each optimization step that are within the decrement bandwidth specified by the end user in step-one (i.e., RMSEA increases up to .001). For example, a .001 increase in RMSEA may represent a .210 increase in the function value. Any estimate positioned on the likelihood function up until this higher fit value is stored.
5. Finally, the end user is then provided with the set of fungible parameter estimates in which they may calculate summary statistics and perform other analysis. For example for θ_1 (e.g., mean = 1.6, minimum = .8, maximum = 1.8). These values then may be compared to the MLE in order to determine the stability of the model parameters.

2.6.5.3 Differences With the Existing FPE Methodology.

The simulated annealing method is both similar to, and differs from, the previous method for calculating fungible parameter estimates (i.e., T. Lee et al., 2017). The method T. Lee et al. (2017) employed for calculating FPEs will be henceforth referred to as the FPE contour approach because all parameter estimates are located along a line (for a two variable example) representing the same value of model fit (i.e., the same value of the discrepancy function). In the FPE contour method, the ML solution is estimated by identifying the maximum (or minimum) of a likelihood function (Figure 2.4). Next, the user identifies a decrement in model fit, and multiple parameter estimates are found along the same level of model fit. While each of these estimates are different, they all represent the function equally well and can be used interchangeably (i.e., they are fungible) (Figure 2.5). Depending on how

different (or similar) these points are from those calculated at the MLE, the researcher can then make a decision regarding their confidence in a particular parameter estimate, and subsequently the overall SEM. T. Lee et al. (2017) recommend using several levels of perturbation as part of any analysis of FPEs. This is an appropriate suggestion for at least two reasons. First, there is not a consensus level of model-fit decrement that should be explored for every model. Exploring a single point limits knowledge of the function to just two sections – the MLE and the FPEs generated at one contour (i.e., a "slice" of the likelihood function in a two variable example) representing a particular decrement in model fit. In addition, calculating FPEs at multiple decrements allows for a better understanding of the function surface. Whereas a simple monotonically increasing function may be reasonably well understood by one (or two) contour(s), this is not the case for a more complicated function. By selecting multiple points, one does not need to make assumptions regarding the underlying functions monotonicity. Thus, it is not necessarily important that the parameter estimates are fungible but rather that the likelihood surface is explored so that researchers may make judgments regarding the level of support for the MLE over other potential parameter estimates. Because the goal is an assessment of parameter estimate stability, the method of generating FPE by function exploration via simulated annealing (or similar algorithm) is better suited than assuring that every estimate is strictly fungible with every other as with the FPE contour method. Thus, the proposed method specifies exploration of the likelihood function without the restriction (i.e., or purpose) of mandating that potential estimates are strictly fungible with one another (see Figures 2.6 and 2.7 for comparison).

2.6.5.4 Guidelines for Use and Interpretation. The level of model misfit induced—and thus the area of the function explored—is a subjective decision. However, it is of critical importance that the decrement in model fit represents a level of model-fit that is not meaningfully different from the level of fit at the MLE. Model fit is intended to

measure how well the data is summarized by a particular model and the resulting parameter estimates. As mentioned, if marginally worse model-fit produces vastly different parameter estimates, it calls into question whether the specific parameter estimate is important or necessary for describing the data. Following Lee and colleagues' (2017) suggestion, decrements of .005 and .01 RMSEA represent appropriate starting points for parameter exploration. Once values are generated from the `psindex` (or by a traditional FPE contour methodology), they must be compared to the MLE. What constitutes a large difference between the MLE and other calculated estimates should depend on the substantive expertise of the researcher. For instance, does the difference between $\theta_{MLE} = .5$ and $\theta_{RMSEA+.01} = .7$ represent cause for concern? While the role substantive knowledge should be important in guiding our interpretation of FPEs, some general guidelines as to what values might be expected in differing analytic conditions is needed.

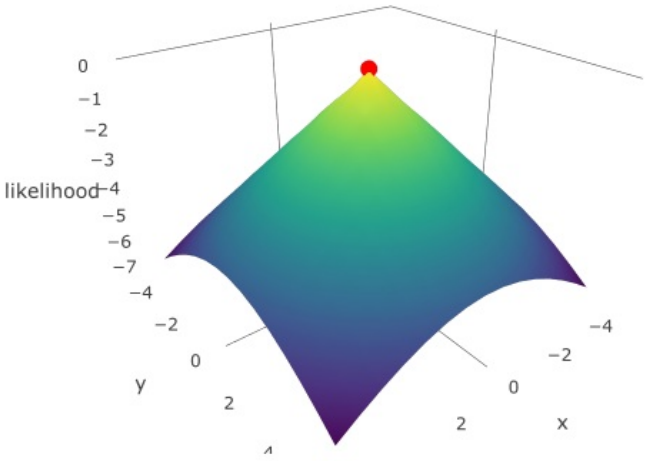

Figure 2.4. Example of a likelihood function that an SEM program maximizes (or minimizes). The red point indicates the optimum and the MLE. Here the MLE for x and y is equal to zero

Illustrative Empirical Example One. This section describes two different real data examples that illustrate how fungible parameter estimates can be interpreted in

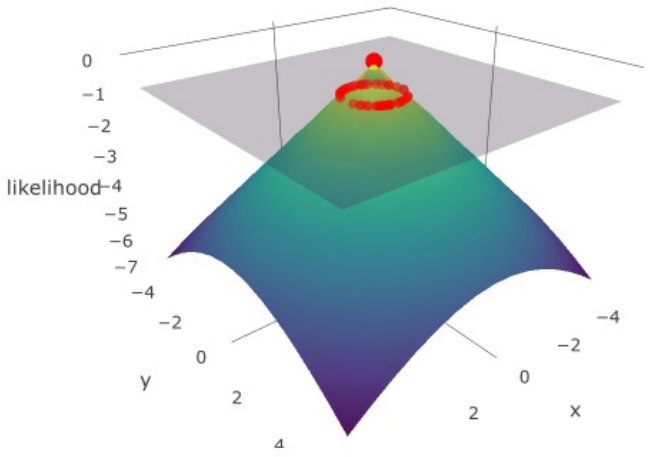

Figure 2.5. Example of a likelihood function that an SEM program maximizes the red point indicates the optimum and the MLE. The transparent plane here represents a specific user-identified decrement in model fit. The FPE contour method then calculates points along this contour line (i.e., the intersection of the function and the transparent plane). Each of these points along the contour line equally represents the function and is therefore fungible with any other point along this plane

practice. For the first empirical data example we use the political democracy dataset ($n = 75$) referenced in Bollen (1989, p. 324). A schematic of the structural equation model to be estimated can be viewed in Figure 2.8. This model was estimated using the `psindex` package in R using the following code:

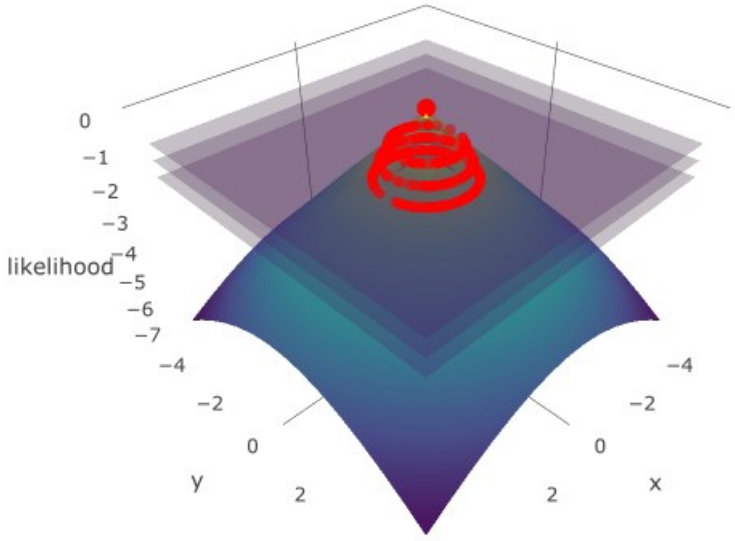

Figure 2.6. Example of the FPE contour method with multiple contours representing different levels of model misfit (i.e., not the MLE). Parameter estimates are calculated at each level and represent the function equally well (i.e., are fungible).

```
install.packages("devtools")
library(devtools)

devtools::install_github(repo = "nietsnel/psindex")

library(psindex)

model_original <- '
# measurement model
ind60  =~ x1 + x2 + x3
dem60 =~ y1 + y2 + y3 + y4
dem65 =~ y5 + y6 + y7 + y8
# regressions
dem60 ~ ind60
dem65 ~ ind60  + dem60
# residual correlations
y1 ~~ y5
y2 ~~ y4 + y6
y3 ~~ y7
y4 ~~ y8
y6 ~~ y8
```

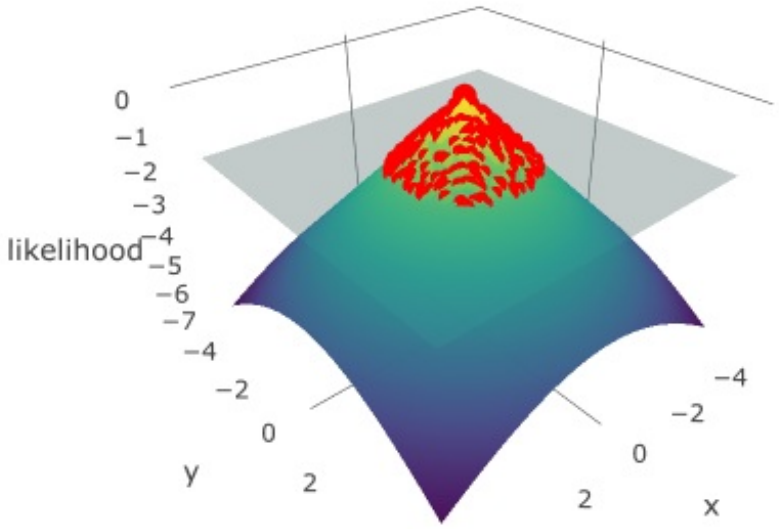

Figure 2.7. Example of a function explored via the simulated annealing algorithm. Here a single maximum misfit boundry is chosen and points between it and the MLE are identified.

```
'
ps_index(model = model_original, data_set = PoliticalDemocracy,
         RMSEA_pert = .005,
         plot_fpe = TRUE, output_long = FALSE,
         frac_plot = .3, iterations_bin = 100000,
         control_genSA = list(threshold.stop = 1e-13,
         max.time = 600))
```

By default, the `psindex` saves the MLE and all parameter estimates within the RMSEA perturbation limit. While a visual examination of the likelihood shape from the most elementary examples (Figures 2.2 and 2.3) can be informative, its not practical in most circumstances. It is difficult to visualize the shape of a complicated function with more than three dimensions. In order to assist with interpretation the `psindex` can instead generate a plot of the FPEs alongside the MLE (Figure 2.9 and Figure 2.10). The output in these two figures indicates that the parameter estimates are moderately stable around most of the parameter estimates and unstable around the variance estimates. In

this model, variance estimates are likely not the focus of the research question and the results should serve to increase the researchers overall confidence in subsequently interpreting the MLE for the parameters of interest. For example, the FPEs generated for the "dem65_dem60" parameter are tightly clustered around the MLE ($MLE = 0.837$, $MLE_{RMSEA+.005} = 0.786$, a 6.09% change from the MLE) and can be considered highly stable. Whereas the FPEs generated for the "dem60_ind60" parameter are more widely dispersed around the MLE ($MLE = 1.483$, $MLE_{RMSEA+.005} = 1.167$, a 21.29% change from the MLE) and might be considered as moderately stable. The data model-fit of this original model at the MLE is RMSEA = 0.035. In summary, our FPE analysis indicates when the upper bound of data model fit using RMSEA is set to RMSEA = 0.045, the parameter estimates are relatively stable suggesting relatively strong support for the MLE.

FPE analysis is not strictly a likelihoodist technique, however. The likelihood principle indicates that all information relevant to inference of θ is contained within the likelihood. This method is dependent on RMSEA which is dependent on sample size to determine the upper limit of the likelihood function that is explored.

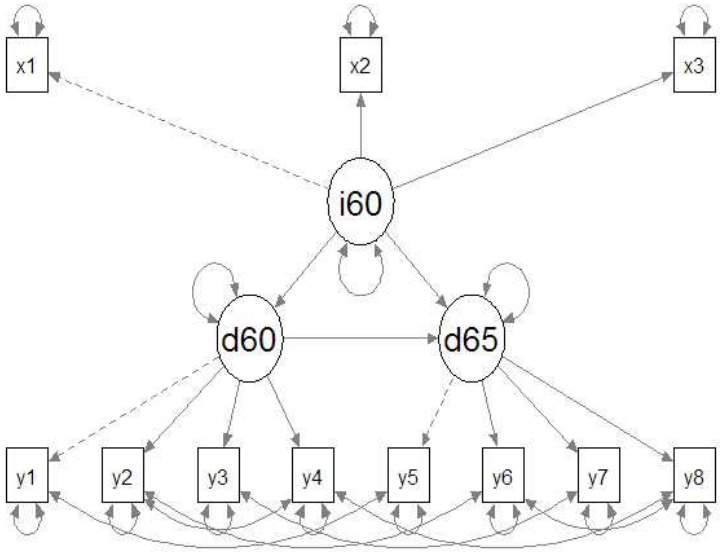

Figure 2.8. Structural Equation Model of the Political Democracy dataset

Illustrative Empirical Example Two. For the second example, we examined the same Political Democracy dataset but estimated the same model as in example one but with a larger decrement in model fit. The overall fit at the MLE is RMSEA = 0.035. Here again, the `psindex` function is used to estimate FPEs but instead this time with a larger decrement in model fit up to +.01 RMSEA (i.e., RMSEA = .045). This model would still be considered by many to be acceptably fitting and the FPEs are, as in the first case, only slightly worse fitting than the original model. However, it is clear that the FPEs show large ranges for several parameter estimates (e.g., dem60_dem60, dem60_ind60). The FPEs generated for the "dem65_dem60" parameter are tightly clustered around the MLE ($MLE = 0.837$, $MLE_{RMSEA+.01} = 0.904$, a 7.99% change from the MLE) and should still be considered as highly stable. Notably, the FPE for the latent path dem60_ind60 show a large range ($MLE = 1.483$, $MLE_{RMSEA+.01} = 2.006$, a 35.29% change from the MLE) and should be interpreted with caution. In this case, inferences regarding the dem60_ind60 parameter should be tempered by the acknowledgement that the MLE estimate is relatively unstable with values of 1.483 and 2.006 describing the data similarly well.

This chapter summarizes the relevant research related to the current state of FPE analysis research. It begins with a discussion of the different types of statistical evidence, and how FPE research fits within a larger set of sensitivity analysis techniques. Next this chapter reviewed the current state of FPE research, why FPE analysis is theoretically justified and how the `psindex` package overcomes previous limitations. Finally, this chapter ends with a real data example that allows readers the opportunity to view an applied example with estimates calculated by the new `psindex` package, and an example of how those estimates might be interpreted. While the uncertainty uncovered by FPE analysis appears to be informative, little research has been done to establish the conditions that cause this uncertainty. The following chapter explains how the PSINDEX along with a simulation study are used to help better understand the causes of unstable parameter estimates.

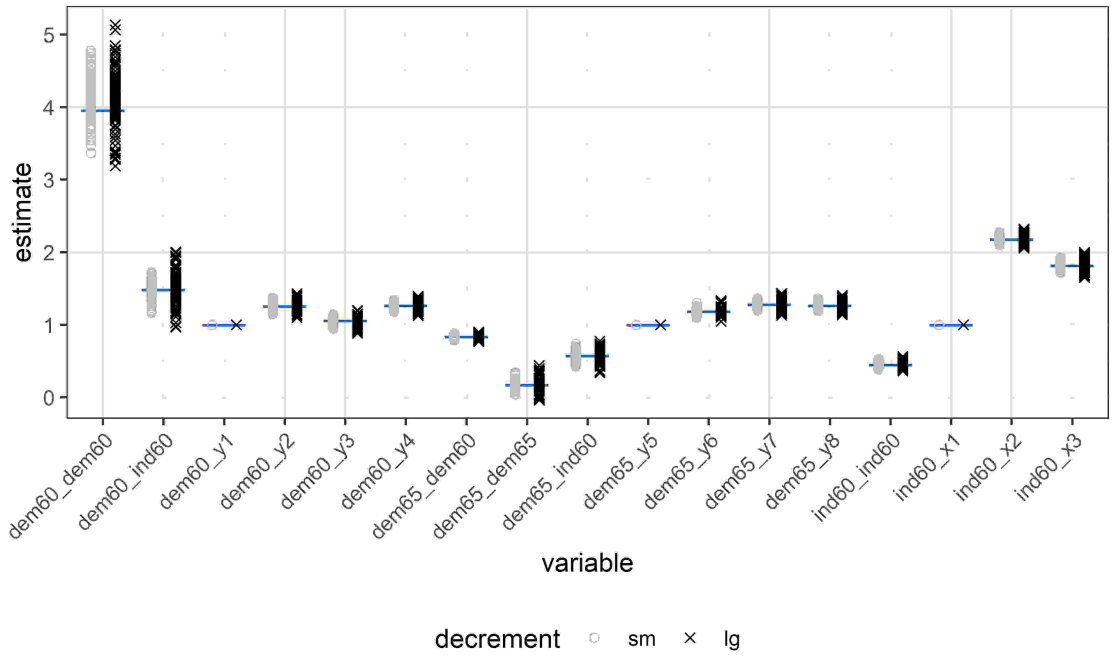

Figure 2.9. psindex Political Democracy FPE output (1/2): The $x-axis$ represents model parameters (e.g., dem60_ind60 is the latent path between the dem60 and ind60 constructs). The $y-axis$ represents the estimated value of the respective parameter estimate. The MLE is is highlighted as a blue line. The FPEs generated for the first model are within a .005 RMSEA decrement range are shown in grey, while those from the second model are within a .01 RMSEA decrement range and are shown in black. We can see here that the points are largely clustered closely around the MLE increasing our confidence in interpreting the MLEs. This figure shows the first half of the model parameter estimates for the example models.

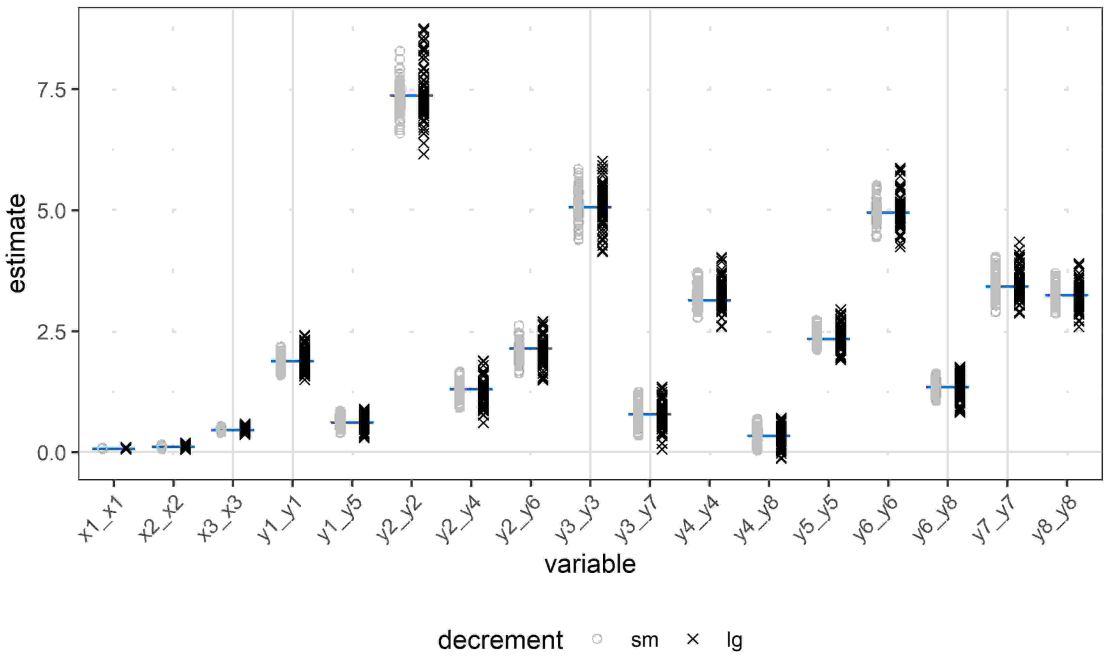

Figure 2.10. psindex Political Democracy FPE output (2/2): The $x-axis$ represents model parameters (e.g., y2_y2 represents the variance of the y2 variable). The $y-axis$ represents the estimated value of the respective parameter estimate. The MLE is is highlighted as a blue line. The FPEs generated for the first model are within a .005 RMSEA decrement range are shown in grey, while those from the second model are within a .01 RMSEA decrement range and are shown in black. We can see here that the points are largely clustered closely around the MLE increasing our confidence in interpreting the MLEs. This figure shows the first half of the model parameter estimates for the example models.

Chapter 3: Methods

FPE analysis has been shown to represent a promising method of quantifying uncertainty in SEM. However, little research has been conducted in order to verify the usefulness of this relatively new methodology. Therefore, a simulation study is proposed in order to conduct FPE analysis under a variety of circumstances in which the researcher might expect uncertain results. This work is, however, to an extent exploratory in that limited past empirical simulation work has been conducted. A simulation study is proposed to facilitate the understanding of FPE behavior especially in scenarios that are analogous to those that might be encountered in empirical research. The aim for selecting representative conditions is also tempered by the need to include conditions that might be less common in practice (e.g., comparing FPE indexes across a large range of sample sizes) but are aimed at clarifying the relationship between parameter uncertainty and manipulated study factors which may not be as apparent when factor levels are indeed too similar. That is, there are two competing goals. The study must include factors that are differentiated enough that their effect on parameter stability (i.e, the range of FPEs) would be measurable if the factor in truth does have an effect on parameter stability. Whereas, conditions similar to real world empirical research must also be included so that plausible arguments for the study's external validity to those conditions can be justified. This

external validity is needed for those applied researchers who may be interested in investigating parameter uncertainty in their own work.

It is possible that the examples that appear in the previous methodological literature are not representative of the information that might be gathered from an FPE analysis. Outwardly, useful examples of FPE analysis may have been chosen inadvertently precisely because of their appearance of providing useful information or by chance. To this aim, this thesis is in line with Popper's falsificationist perspective (Popper, 1959). By conducting a more comprehensive simulation study, the FPE analysis framework may be strengthened by not being refuted. Refutation is defined as not providing useful information regarding parameter uncertainty when known simulation conditions would show otherwise.

3.1 Research Questions

This dissertation aims to test whether FPE analysis might provide useful information in an analytical context. Thus, two questions are examined: 1) Which conditions result in the largest FPE ranges? 2) How do results in FPE analysis differ when defined by AIC rather than RMSEA? AIC indexed FPEs are thought to capture epistemic and alleatory uncertainty whereas RMSEA are expected to index primarily epistemic uncertainty. Thus, it is hypothesized that the FPE range will be more sensitive to sample size when defined by AIC than RMSEA. Conversely, it is not expected that the magnitude of AIC indexed FPEs will differ only in their magnitude from RMSEA indexed FPEs.

Aiming to resolve these two questions is helpful for two reasons. First, by understanding which conditions result in the largest FPE ranges, those undertaking FPE analysis in their own research can better understand what conditions might be responsible for either small or large FPE ranges and thus may gain insight into ones statistical model

and results. Secondly, by considering how the definition of FPE indexes affects the type of uncertainty measured by FPE analysis the decision to select one index in lieu of another can be more thoughtfully based on the analytic need. Diagnosing results based on FPE analysis is not possible without first understanding the information conveyed in FPE analysis. This research aims to build on past work to establish this understanding.

3.2 Simulation Design

A Monte Carlo simulation study is proposed in order to investigate the usefulness of the FPE analysis framework. Data generation for all manipulated conditions was conducted using R (R Core Team, 2018). The Parameter Stability Index (`psindex`) package in R was used to calculate fungible parameter estimates for all models. This package utilizes a simulated annealing algorithm and was purposely created in order to facilitate both practitioner use and methodological investigation into fungible parameter estimates (Prendez & Harring, 2019). In order to complete this research study within a practical time frame the study will be conducted using parallel computation. Previous methods for calculating FPEs were not easily accessible whereas the `psindex` package presents the potential for studying FPE analysis for the first time in a relatively large number of modeling scenarios.

3.2.1 Measuring Uncertainty – Outcome Measures

In order to investigate the utility of the FPE methodology, the range of FPEs will be investigated. Fungible parameter estimate analysis, however, generate a large amount of data by means of the thousands of fungible estimates for every estimated model parameter (this can be seen in Figures 2.9 and 2.10). For each parameter, the percent change between

the MLE estimate and the most distal FPE will be calculated. This method is useful because it allows for summary of the FPE analysis to a single interpretable value for each output variable. Small percent changes represent low levels of uncertainty whereas larger percent changes represent relatively higher levels of uncertainty. Secondly, this method is justifiable because the most extreme values are the most relevant. In an FPE analysis no distributional argument is made from the resulting FPEs. This further justifies the use of the simple and interpretable MLE percent change measure. The raw FPE range is also included as a measure of parameter uncertainty. This measure is defined as the raw range from the two most distall fungible parameter estimates.

3.2.2 The Utility of Monte Carlo Simulations for Studying FPE Analysis

FPE analysis is a method for estimating the peakedness of the likelihood function in n-dimensional space. A relatively flat distribution represents an uncertain estimate (i.e., many estimates describe the data similarly well) and a peaked likelihood function represents a parameter estimate with relatively lower levels of uncertainty. Therefore, the purpose of the research is to understand how the shape of the likelihood is affected by different factors. While each dataset will result in a slightly different MLE (from which FPEs are calculated) the relevant quantity is the expected value across simulations. The expected value of a given model parameter of interest, $\hat{\Theta}$, in a Monte Carlo simulation is determined by calculating the mean of the parameter estimates across R replicates shown below in Equation 3.1.

$$E\{\hat{\Theta}\} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r \tag{3.1}$$

The true value of parameters are considered as fixed quantitites in frequentist statistics. While the true value of the the parameter of interest are not considered random variables the estimands are because the value of the estimate is generated according to a probability distribution.

Monte Carlo Simulation studies are useful in part because they allow for calculation of the expected MLE based on sample values (even though there is variability due to the random process in which data are generated). When each data set is simulated it has an accompanying likelihood function that also similarly differs between samples. While the shape of this function has a random component, it too can be characterized by calculating its expected value over multiple instantiations. This process can be observed more clearly using a simple example involving an individual parameter estimate and its requisite likelihood function.

Figure 3.1 shows the MLE and a section of the likelihood function from which it was calculated. This example (and code) was adapted from a similar example by Goldfeld (2017) which illustrates how Monte Carlo simulations can be used to investigate the factors affecting FPEs.

Figure 3.2 displays MLEs calculated from multiple simulated datasets as well as their accompanying likelihood functions. The expected value of the MLEs is close to the simulation target value (shown in the left sub-figure). Each MLE is a summary of a likelihood function. The accompanying likelihoods of these same MLEs are shown on the right sub-figure in 3.2. While each likelihood function is slightly different, this variation can be summarized over repeated simulated iterations. By using this summarized information, it is possible to study the effect of other factors and not the random process used to

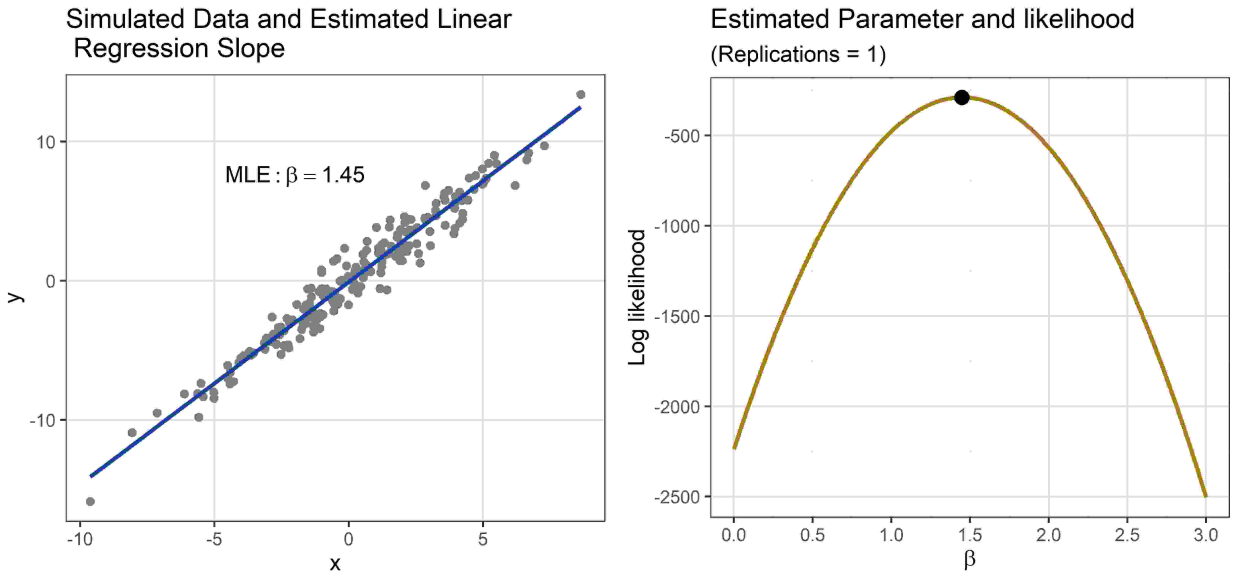

Figure 3.1. In the left hand panel a simulated dataset and fitted linear regression slope, generated from a population parameter target slope of 1.5. In the right hand panel the accompanying likelihood function. The black point represents the maximized likelihood parameter estimate of $\beta = 1.45$.

generate the data-set. That is, the average characteristics of the likelihood function can be summarized in terms of how they are affected by the independent variables of the simulation study (e.g., sample size, model specification).

Fungible parameter estimate analysis is implicitly a comparison between those fungible estimates and the MLE. The distance from the MLE to the FPE is a measure of the support for the MLE relative to other candidate values (in 2D it is a descriptor of the flatness or relative peakedness of the likelihood function). In practice this distance can be measured in several different ways, by calculating the raw or percentage change from the MLE to the most distal FPE or by calculating the total range between the most distal FPEs, for example.

Figure 3.3 serves as a graphical illustration of how a Monte Carlo study can be used to investigate likelihood functions under different study conditions. This figure exhibits MLEs,

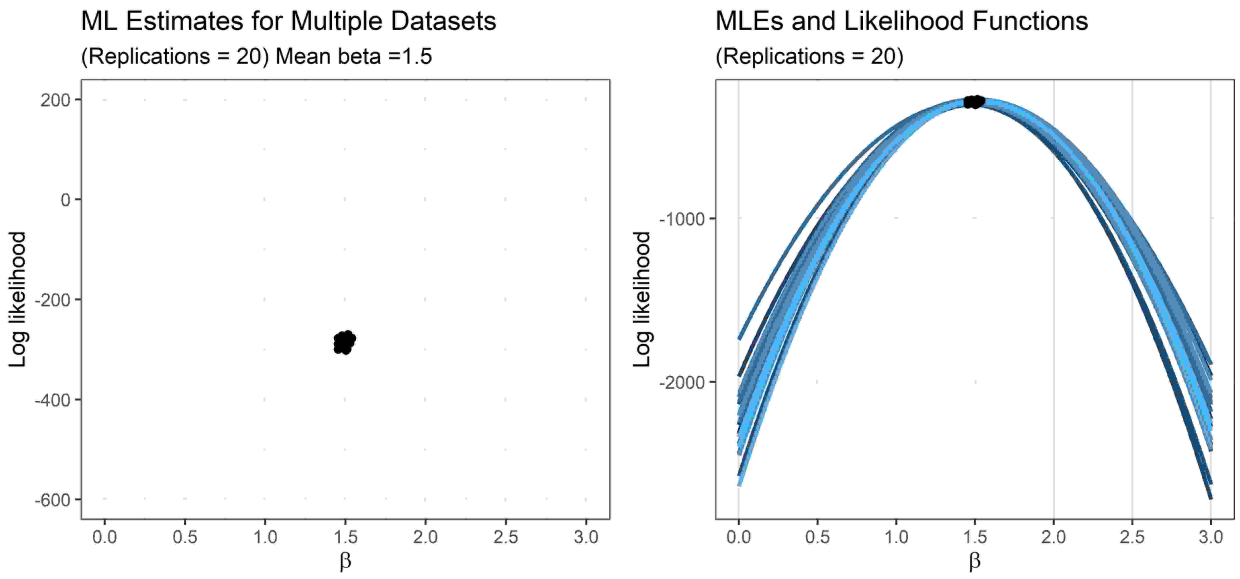

Figure 3.2. Left: Maximum likelihood estimates based on repeated simulated datasets. These datasets are repeated instantiations of the data generated in Figure 3.1 (i.e., simulation population parameter target of $\beta = 1.5$). Right: The accompanying set of likelihood functions that were maximized in order to reach the MLEs shown on the left.

and a section of their accompanying likelihood functions generated for two different data sets (i.e., conditions A, and B). This figure demonstrates that the underlying likelihood may differ between datasets despite similar or equal MLEs. Here two sets of likelihoods bisect a user defined fit-decrement at two different points. In this 1-variable example there are only two fungible estimates per condition. Although each replication leads to a different likelihood function the expected value of the likelihood function can be calculated over repeated replications. The mean value of FPEs is then represented in Figure 3.3 by four squares. The figure exhibits differing FPE ranges per condition. Condition B displays a larger raw-range, $\{0.787, 2.240\}$, between the two FPEs when compared to condition A, $0.411, 2.63$. Despite the variability in the likelihoods per replication this example demonstrates that there is less uncertainty in condition A than condition B.

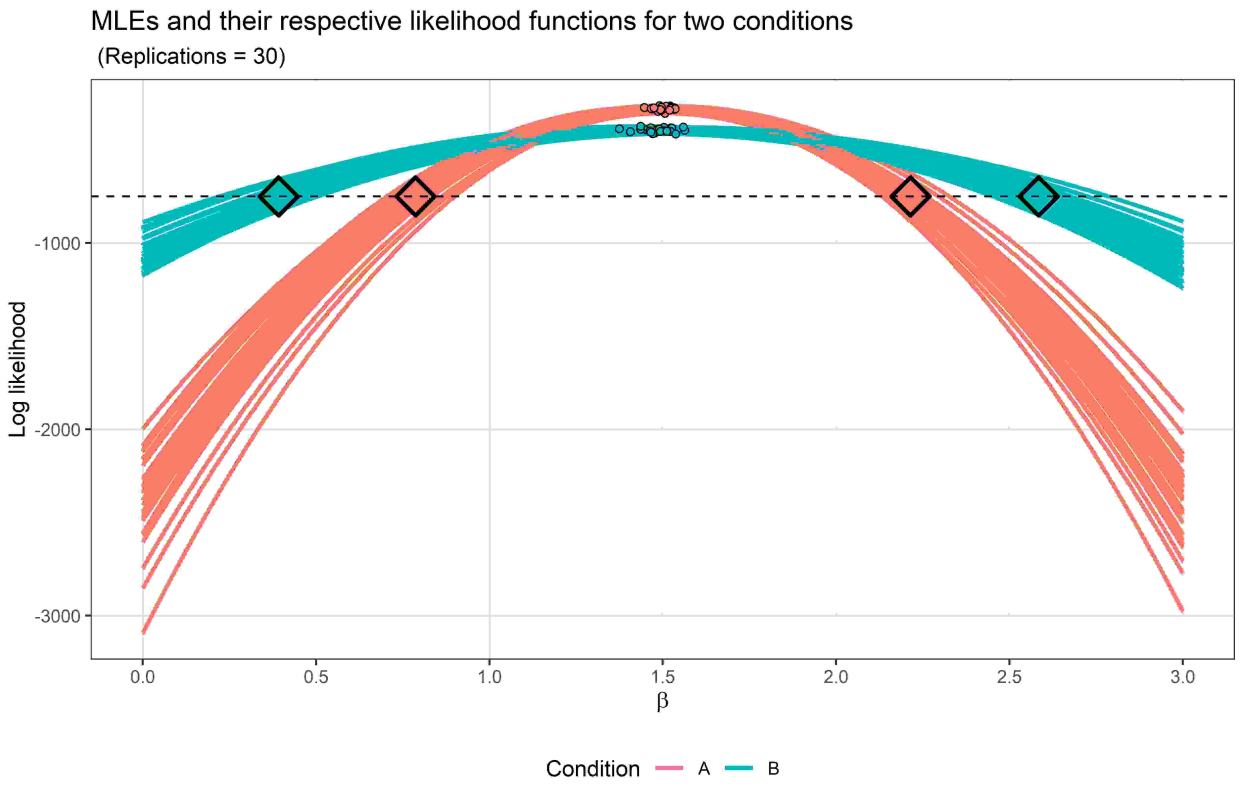

Figure 3.3. MLEs and their respective likelihood functions for two simulated datasets replicated 30 times per respective dataset. The MLEs for both datasets are represented by the small set of circles centered around $\beta = 1.5$ (x-axis). The expected value of these two sets of MLEs is the same but the shape of their accompanying likelihood function differs. The dashed horizontal line represents a user-defined decrement in model fit in which fungible parameter estimates are calculated. The four squares represent the mean value of the fungible parameter estimates calculated at this level of model fit decrement across replications.

3.2.3 Data Generation

Data were generated using the following method:

1. Data were simulated according to the conditions outlined in this chapter using the `simsem` package in R. Using this package, a total of six distinct datasets were simulated based on large sample sizes (n = 5,000,000).

2. The covariance matrices of these six datasets were then saved and used as input for the `mvrnorm` function in R. The matrices were confirmed as valid covariance matrices

using the following two steps. First, matrices were inspected to verify that they were symmetric (i.e., $A_{XX}^T = A_{XX}$). Secondly, matrices were tested if they were positive definite by inspecting that all eigenvalues were positive. Directly utilizing covariance matrices and the `mvrnorm` function within the simulation provided a compact and fast method of simulating data. Data generated by these correlation matrices and `mvrnorm` were then tested in the first pilot study to ensure that the data were simulated correctly.

3.2.3.1 Procedure for Non-Convergence. Statistical point estimators aim to achieve a unique solution. As mentioned previously, for maximum likelihood estimators model convergence is obtained when the global minimum of the fit function is located. This optimal value, however, cannot always be identified. Small data sets (Paxton, Curran, Bollen, Kirby, & Chen, 2001a), or relatively complex models (e.g., high-dimensionality, or mixed models), are less likely to converge. In Monte Carlo simulation studies a subset of replicates might not converge despite being generated and estimated using the same parameters. When replicates do not converge, they must then be removed from the analysis or otherwise modify the study (e.g., change convergence criteria). The decision to replace these replicates or adjust the simulation design may change the interpretation of the studies results. Replacing replicates with additional iterations allows for the level of planned precision to be maintained. The decision to not replace replicates may, however, limit the generalizability of the results to only those conditions and replicates that successfully converged.

For this study the decision was made to replace non-converged iterations with new replicates. In addition, the smallest sample size was also increased from N = 100 to N = 200. This decision reduced the rate of non-convergence. For the "misspecified 1" condition,

replicated 30 times, (discussed in the Manipulated factors section) and a sample size of 100 there were a total of 7 MLE, and 8 FPE non-convergences. For the same condition with a sample size of 200, there was 1 MLE, and 1 FPE non-convergence. For this study, discarding a data set and re-running replicates occured under either of these two distinct types of non-convergence. First, if the initial maximum likelihood estimation did not converge a new data set was generated (ML non-convergence). Secondly, a new data set was also generated for any replication in which the simulated annealing algorithm did not locate any fungible parameter estimates (FPE non-convergence). This process was repeated until the minimum number of replications was achieved. As mentioned previously, the simulated annealing algorithm is not deterministic. Increasing the allowable iterations of the algorithm will eventually allow the `psindex` module to successfully identify FPEs. Thus, increasing the number of iterations is one method of locating FPEs rather than reiterate the current replication. Eventually, the number of iterations was chosen in order to achieve a balance between long algorithm run time and iteration restarts.

3.2.4 Assessing Fidelity and Variance of Data Generation procedure

This section evaluates two different aspects of the data generation process. First, the fidelity of the simulation is assessed at the population level, and second, the variability of the data generation procedure is measured.

The first section demonstrates the degree to which the data generation procedure is able to generate the data in accordance with the factors and levels set forth in the study design. This will be quantified by using single large scale replicate (i.e., 50 million). It is useful to consider how closely the data are generated in accordance with their simulation

targets. If the data generation procedure cannot achieve the target levels it may affect the interpretation of the simulation results.

The second aspect is the variability of the data generation procedure and the estimation of Monte Carlo error. Monte Carlo error is the degree of between-simulation study variability that can be expected. Understanding this value helps to justify the number of replications that should be undertaken. This pilot study was conducted using a sample size of 1000. This represents the middle sample size condition out of the three sizes used (200, and 5000 are the others). The advantage of selecting the middle sample size condition is that the variability based on these results should also be middling compared to the higher variability, and lower variability of the smaller and larger sample sizes, respectively. Basing the number of replications on the smallest sample size is more conservative decision but is also not representative of the variability for the other larger conditions whereas choosing the middle condition is a compromise between the optimistic and pessimistic sample size condition.

Simulation Result	dem65-ind60	dem65-dem60	y7-y11	y3-y11	ind60-y1
Simulation target value	0.182	0.7	0.11	0.18	0.92
Estimate	-0.164	0.773	0.111	0.181	0.92

Replication = 1

Table 3.1. *Simulated Data Fidelity (N = 50 Million)*

Note. See Appendix A1 for all parameters.

3.2.4.1 Simulated Population-Level Data Fidelity.

First, the fidelity of the simulation to the targeted parameters values is examined. Correct data generation was assessed by confirming that parameter estimates based on pilot data were not far from the population parameter target values used to generate the data. The single large simulation of N = 50 Million was used to minimize the role of random chance. The data were generated and estimated according to the model shown in Figure 3.4. A subset of these parameters is shown in Table 3.1. One parameter value, dem65-ind60, showed a notable departure from its simulation target. The estimated value displays a negative relationship and is approximately 11% greater in magnitude than the target parameter. The model used in this study is inspired by the real-data analysis from Bollen, 1989. However, the model used in the current analysis has an additional path not present in this model. The remainder of the estimated simulation results are not far from the population parameter target values used to generate the data. Overall, these results indicate that the data is being generated as expected. The full set of parameters from this condition are shown in Appendix A1.

3.2.4.2 Sample Variability and Monte Carlo Error.

This dissertation project utilizes a Monte Carlo simulation in order to investigate the behavior and thus the utility of an FPE analysis. This analysis takes place in multiple steps. First, sample data is generated according to a population model and a given study condition. Second, FPEs are estimated based on these generated datasets. Previous research has primarily undertaken only the second step (i.e., generating FPEs on applied datasets). However, the benefit of a simulation study is that it provides methodologists with insight into the behavior of a technique given known population parameters. This is especially important when the

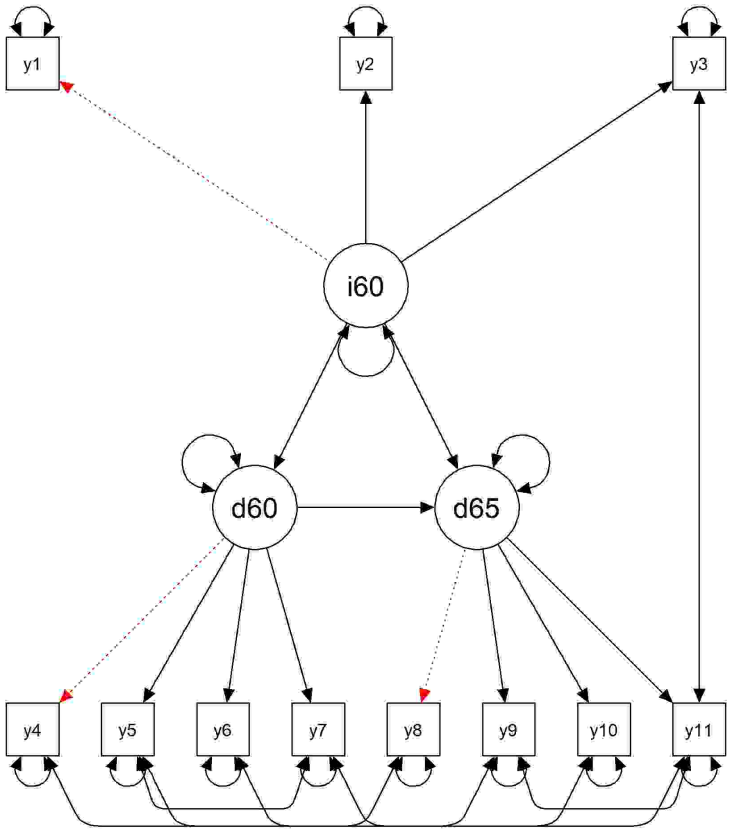

Figure 3.4. Pilot Model. Black arrows represent freely estimated parameters. Red dotted arrows represent fixed parameters

behavior is complicated, or unknown in particular circumstances (i.e., small samples) (Paxton, Curran, Bollen, Kirby, & Chen, 2001b).

However, this often elucidating technique is not determinstic as random numbers are used. Because of this care must be taken to understand how differences in each generated data set could effect summary statistics and our resulting interpretation. Importantly, the sample size must be large enough so that the mean value of these parameters is close to the target simulation parameters. The summary statistics based on repeated samples in a Monte Carlo simulation study are also subject to variability. That is, if the simulation

study were to be computed again the results would differ. This between simulation study variability is called Monte Carlo Error (MCE). This is defined below in 3.2.

$$MCE\{\hat{\Theta}\} = \sqrt{var[\hat{\Theta}_{sim-1,\ldots,Sim-N}]}. \tag{3.2}$$

Here $\hat{\Theta}$ is the parameter of interest that is calculated according to 3.1. The value of this parameter for multiple simulation studies is represented by the radicand vector (Sim-1, Sim-2,...Sim-N).

Monte Carlo error must be mitigated so that the results, though based on finite samples and replications, are meaningfully representative of the phenomenon under investigation. In order to reduce this error due to chance, a sufficient number of replications must be generated. There are several methods that can be used to estimate MCE. For instance, it is possible to replicate the simulation study multiple times and calculate the variability of the results directly. Depending on the simulation study this can be prohibitively time consuming so fortunately this is not necessary. Instead, the between study variability can be conceptualized as the simulation standard error:

$$SE_{SIM} = \frac{s}{\sqrt{R}}. \tag{3.3}$$

Here s is the standard deviation of a given parameter of interest, $\hat{\theta}$, across all within-simulation replications (R) in a Monte Carlo simulation study. The standard error of the simulation is an estimate of the standard deviation for the population of potential simulation studies.

Monte Carlo error example. A simple simulation study can be used to illustrate MCE. This example is similar to the model and example used by Koehler, Brown, and Haneuse (2009). This simulation example is used to compare 3.2 and 3.3. The model relates a binary outcome variable to two parameters: logit $P(Y = 1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. The values of β_1 and β_2 are -1 and log(2), respectively. The simulation study had a sample size of n = 100, and 15,000 replications. The observed mean value of $\beta_2 = .734$. The entire simulation study was then subsequently repeated 200 times. The results of this example are featured in Figure 3.5.

Figure 3.5 displays the variability within-simulation-study estimates (i.e., variability between replications), and the smaller, but still present, between-simulation study variability shown on the right. The variability shown on the left is the same variability discussed previously (see discussion regarding in Figures 3.1, 3.2, and, 3.3) while the variability on the right side represents Monte Carlo Error. Because we have re-run the simulation study many times we can estimate the MCE error directly using 3.2 (i.e., the standard deviation of the values represented on the right side of Figure 3.5). Alternatively the MCE can be estimated using the SE_{SIM} definition (3.3) which estimates the MCE using a single simulation study. The results from the two methods were 2.244e-3 and 2.231e-3, for the MCE, and SE_{SIM} method, respectively. The two numbers are close, matching through the ten-thousandths place. Therefore it is reasonable choice to use the SE_{SIM} method as an estimate for MCE for this project. Other methods for calculating MCE have also been suggested by Koehler et al. (2009) which include an integral based, and "Jacknife-after-boostrap" method.

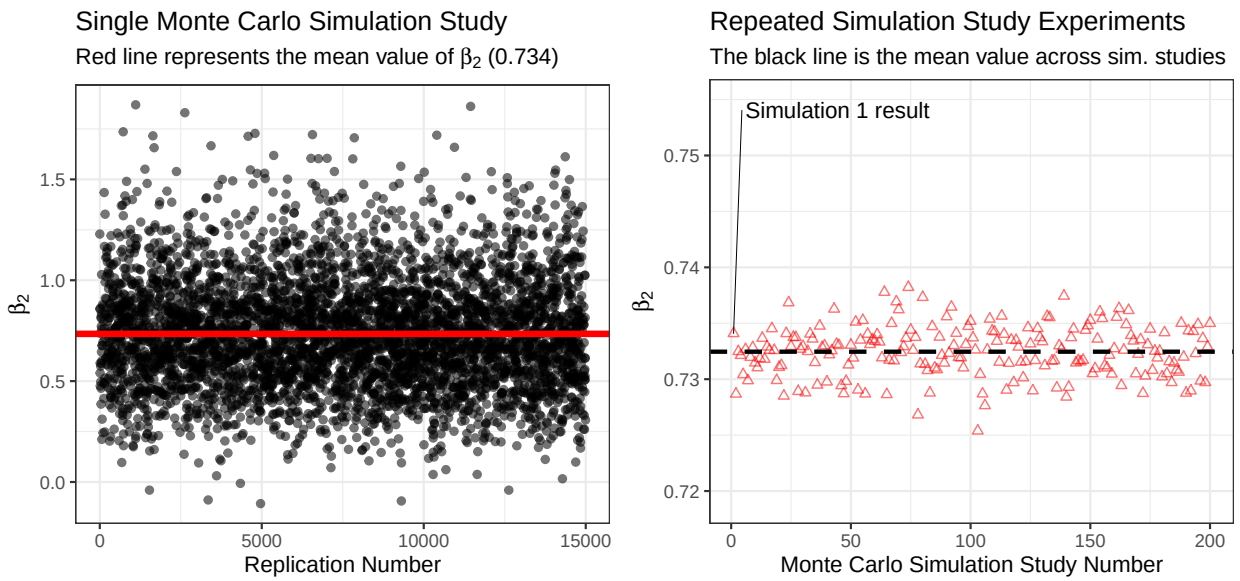

Figure 3.5. Left: Parameter values at each replication of the Monte Carlo Simulation Study. Right: Summary statistics for the same simulation study run repeatedly. The annotated point represents the summary (mean parameter value) of the first simulation study shown on the left.

Analogies to traditional experiments. In a traditional experiment (non-simulation study), a random sample is drawn from a larger population. Imagine that parameter estimates for means and standard errors are calculated for this experiment. The standard errors in this experiment are inferential statistics that estimate the variability of all potential sample means with similar characteristics (i.e., same variability, and sample size) around the population mean based on the Central Limit Theorem. These would be analogous to the individual points on the left side of Figure 3.5. While the points on the right side of this figure might be thought of as a collection of samples or a type of meta-analysis. The distribution of these individual studies represents the uncertainty in any one study mean value. Specifically, how close the study results—the estimated means—are to the true population means (i.e., population of meta-analysis or simulation studies) is the goal in estimating Monte Carlo Error.

3.2.5 Replications and Simulated Annealing Maximum Iterations

A second simulation study was conducted in order to estimate the level of Monte Carlo error in the study outcome measure. This was done in order to select how many samples (i.e., replications) would be needed in order to obtain informative and meaningful results. This pilot was also based on the same model shown in Figure 3.4. In this pilot, the number of replications examined was 30, 60, 90, and 120. Each of these conditions was generated with a sample size of n = 1000. The results of the second pilot study can be seen in Tables 3.2, 3.3, and 3.4. In order to summarize this work all values shown represent the mean values based on all 32 freely estimated parameters in the Pilot Model.

As expected the standard error of the simulation decreases with the increasing number of replications (see Table 3.2). Increasing the maximum allowable SA iterations decreases the SE of the simulation outcome but only for larger replication levels. This relationship is less straightforward than that of increasing the number of replications. Increasing the allowable SA iterations likely allows for locating of more distal outcomes (see Table 3.4). When the number of iterations is relatively low these distal outcomes increase the SE of the outcome measure.

It is useful to consider the relative benefits of increasing replications and or algorithm iterations calls when assigning computational resources. Table 3.3 demonstrates the time required to compute a singular condition under differing replication and maximum iteration levels. The effect of increasing the iteration maximum is a non-linear increase in computational time. Based on this preliminary analysis each condition will utilize a 1000 SA iteration maximum and will be replicated 120 times. This choice represents a compromise between precision and computational time.

Reps	max SA iterations	*M* param. % change MCE	Mdn param. % change MCE
30	1000	1.23	0.76
30	2000	1.95	0.84
30	3000	1.39	0.90
60	1000	1.08	0.56
60	2000	1.24	0.57
60	3000	0.93	0.61
90	1000	5.98	0.46
90	2000	0.97	0.49
90	3000	0.77	0.50
120	1000	0.65	0.41
120	2000	0.73	0.40
120	3000	1.60	0.40

Table 3.2. *Monte Carlo Error (Sim. Standard Error) by Replication and Max. SA Iteration.* **Note.* All iterations converged to the MLE

Reps	max SA iterations	Avg. Num. FPEs	Time (minutes)
30	1000	11048	29
30	2000	13536	104
30	3000	14365	177
60	1000	11048	61
60	2000	13536	182
60	3000	14365	303
90	1000	11044	86
90	2000	13534	251
90	3000	14364	412
120	1000	11061	63
120	2000	13543	208
120	3000	14370	323

Table 3.3. *Average Number of FPEs Estimated and Study Time*

3.3 Manipulated Factors

3.3.1 Estimated Models

The base model used for the simulation is a three factor structural equation model inspired by the model used previously in the real-data analysis demonstration of the `psindex` package (originally from Bollen, 1989). Data for this model will be generated according to the model shown in Figure 3.6. This is the same model used in the Pilot Study discussed earlier. The population values for the simulation were based loosely on the

Reps	max SA iterations	*M* FPE % change estimate*
30	1000	9.89
30	2000	13.78
30	3000	13.16
60	1000	11.29
60	2000	13.11
60	3000	12.76
90	1000	16.69
90	2000	13.24
90	3000	12.70
120	1000	11.89
120	2000	11.90
120	3000	13.33

Table 3.4. *FPE Percent Change Estimate by Replication and Max. SA Iteration*
**Note.* This estimate is the sim. outcome averaged across 32 study parameters.

maximum likelihood estimates from the empirical Political Democracy dataset shown previously. The purpose of this was to create a relatively more complicated and realistic dataset. While simulation research often utilizes uniform parameter values (i.e., each loading representing an equally reliable indicator of their respective latent factor), it does not reflect a realistic scenario and may especially be detrimental to understanding FPE analysis.

While any model choice here would be insufficient in representing the large number of potential models in which an FPE analysis might be taken, this model was chosen for its familiarity with readers, and its relative simplicity. Readers will benefit from being able to compare this basic model to other simulation studies which frequently utilize relatively simple structural models. Secondly, the structure of the model was chosen because testing has indicated relatively low problems with ML convergence, estimation of FPEs.

Once data has been generated according to the model in Figure 3.6, one of three different estimation models will be used. The first model, uses the generation model as the

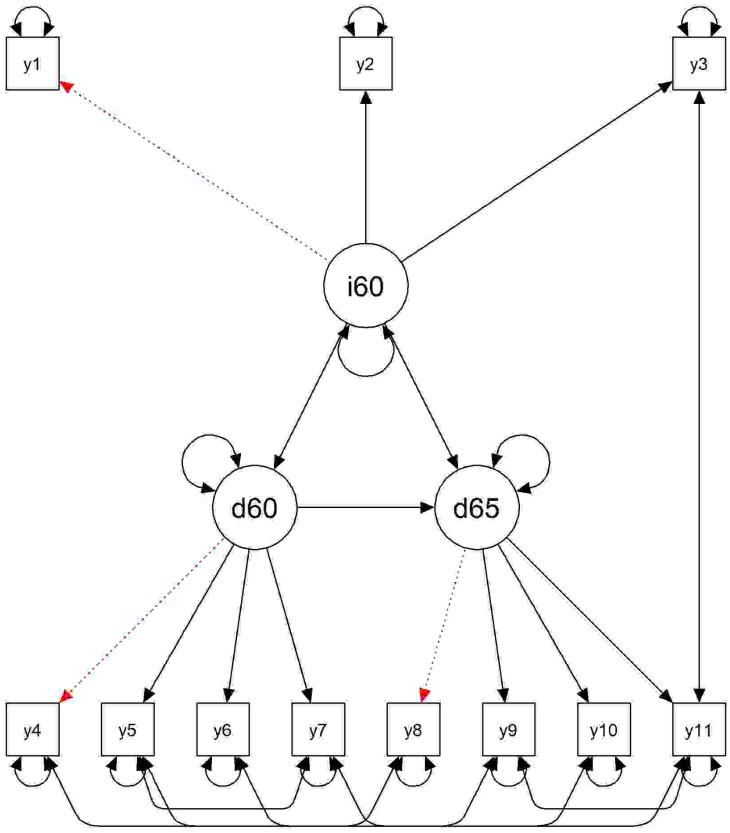

Figure 3.6. Generating Model. Black arrows represent freely estimated parameters. Red dotted arrows represent fixed parameters

estimation model and thus serves as a perfectly fitting base model. In this model RMSEA will be equal to zero. This model is included because it serves as a benchmark for understanding FPE analysis. FPE analysis in this situation should produce relatively small FPE ranges compared to any other misspecified condition. This scenario should provide no evidence of parameter uncertainty. The second and third models both involve model misspecification.

The second model, "misspecification 1," is again similar to the previous empirical example (see Figure 2.8) and can be seen in Figure 3.7. This model was chosen for two

reasons. First, all models in reality are simplifications of truth and are therefore misspecified. This model is misspecified and constitutes a more realistic example of what researchers can expect regarding FPE behavior. This "misspecification 1" model has a moderate level of model misfit —RMSEA = .030.

Model three is displayed in Figure 3.8. This model misspecifies the path between the i60 and d60 latent variables in addition to the path between y7 and y11. The RMSEA for this model for this model is 0.074. Each of the experimental models (see Figures 3.6, 3.7 and 3.8) were simulated with an n = 1,000,000 in order to obtain an RMSEA that approach a population estimate. Previous research has also used relatively relatively simple models to study model misspecification (Fan, Thompson, & Wang, 1999; Saris, Satorra, & van der Veld, 2009).

3.3.2 Model Fit Index

3.3.2.1 RMSEA as a Measure of Model Decrement. RMSEA will be used as a measure of model misfit and as a method of measuring model decrement. RMSEA was included for two reasons. First, RMSEA is one of the most popularly used methods for assessing data-model fit within a SEM framework. Secondly, one of the key assertions underlying FPE analysis is that the estimates are *fungible*. If the estimates cannot be thought of as essentially fungible or interchangeable then the analysis is not useful. Using a metric that practitioners and methodologists are familiar with is important. If there is not an accepted measure of model fit, then it is more difficult to judge how close estimates would need to be to be credibly viewed as representing essentially equally well-fitting estimates (i.e., FPEs).

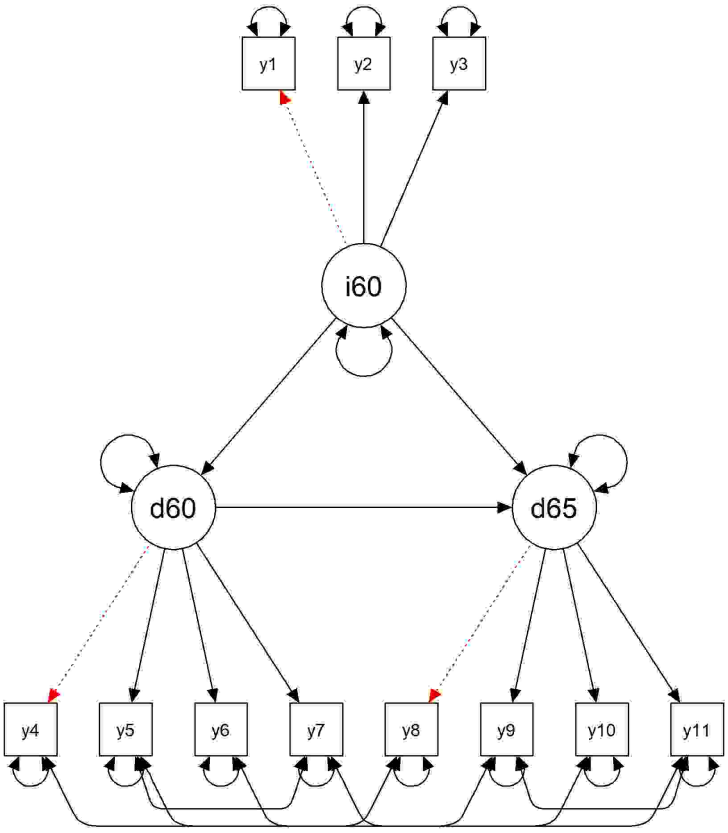

Figure 3.7. Misspecification Model 1. Black arrows represent freely estimated parameters. Red dotted arrows represent parameters fixed to 1

Thus an FPE analysis that uses the familiar metric of RMSEA helps to communicate the results and improves the validity of the analysis. One of the problems with using the raw values of the discrepancy values (or percentages thereof) is that they do not have a familiar basis for interpretation. For the purposes of this study RMSEA perturbations equal to .001 and .01 will be used. RMSEA increase of .001 and .01 represent relatively small changes in data model fit and both values have been used previously in FPE research (Pek & Wu, 2018; T. Lee et al., 2017).

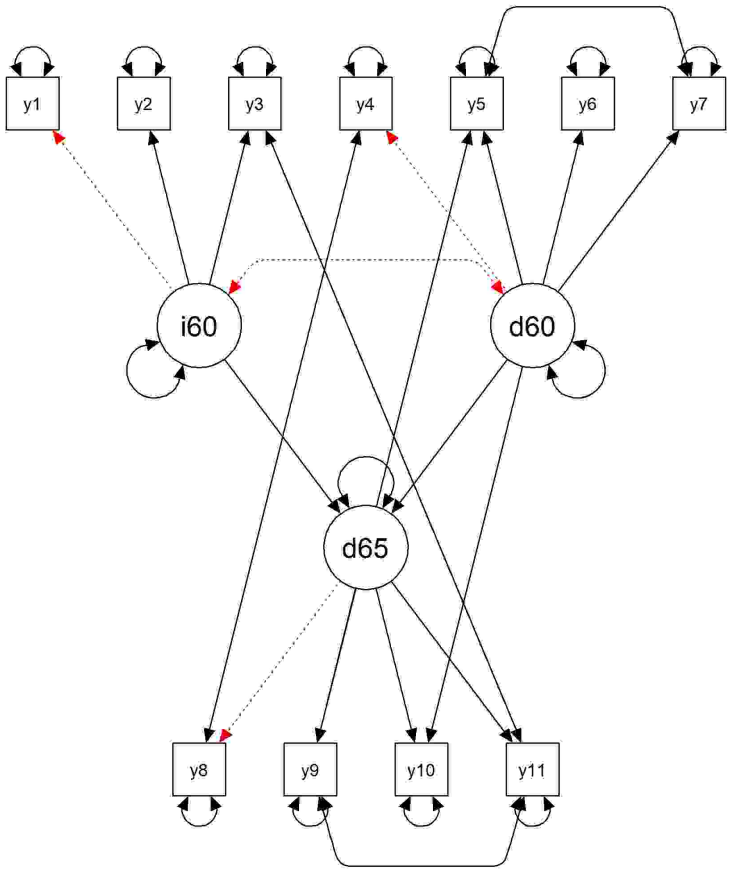

Figure 3.8. Misspecification Model 2. Black arrows represent freely estimated parameters. Red dotted arrows represent fixed parameters. The first latent factor indicators (highlighted with red arrows) are fixed to 1 and the path between the "i60" and "d60" latent factors is fixed to zero.

3.3.2.2 AIC as a Measure of Model Decrement.

AIC will also be used as a second measure of measuring model decrement. AIC is an intuitive choice for this type of measurement because it is frequently used as a measure of model comparison. Rather than comparing between models, as is typical, AIC will be used to adjudicate whether models represented by different parameters estimates (AIC_i) compare against that of the ideal maximum likelihood estimates (AIC_{min}). For the purposes of this experiment, values will be tested under only a single level of model decrement. Any model with an $\Delta_i < 2$ (i.e.,

$\Delta_i = AIC_i - AIC_{min}$) will be considered a FPE. These values were chosen based on the guidelines given by Burnham et al. (2002). AIC values are highly sample size dependent. This feature differs from some other fit measures commonly used in SEM. While this is a desirable feature in many modeling contexts, it is not necessarily useful in an FPE analysis. Sample size will be modified as it is important to understand the behavior of AIC under different sample size conditions in conjunction with other manipulated factors.

Lastly, only models that are reasonably well fitting are included in this analysis. While the second misspecified condition (RMSEA = .074) is above the RMSEA threshold commonly used it was still selected as an important level. Restricting the range of models to only "good fitting" models does not allow for an understanding of whether the range of FPEs increases dramatically as model fit decreases. This desire to examine the behavior of FPEs more broadly is balanced with the notion that this project aims to include conditions that might be further analyzed in practice (rather than including extreme conditions).

3.3.3 Sample Size

Sample size will be manipulated in the simulation and will take on three different levels: 200, 1000, and 5000. These sample size choices were made because they represent large enough values that the study was likely to converge in most circumstances. The lowest sample size condition was initially selected as n = 100. This condition was changed to 200 due to large levels of non-convergence at both the MLE, and FPE estimation stages during the pilot work analyses. Furthermore, there is a reasonably large difference between the small and the largest sample size conditions in order to better understand the effects of sample size in a FPE analysis. AIC is heavily dependent on sample size which should make

the differences in model fit more pronounced at the larger sample size. RMSEA, as discussed earlier, is affected by sample size but to a much smaller extent. Thus, the sample size is expected to affect the range of FPEs differently depending on the index. Using measures that were not sensitive to sample size were initially important in differentiating FPEs from standard errors. As this distinction has been made clear (Pek & Wu, 2018), it is useful to investigate measures both sensitive and non-sensitive to sample size so that valid inferences can be made. It is useful to investigate whether using a measure that *is* sensitive to sample size can be useful when understanding uncertainty from a more holistic perspective. That is because AIC indexed FPEs are thought to capture both epistemic and alleatory uncertainty whereas RMSEA are expected to index primarily epistemic uncertainty. While it may be appropriate to isolate the uncertainty due to the modelling uncertainty (i.e., epistemic) it may be useful to contrast this to a measure that captures both sources. Thus, a measure that includes both sources of uncertainty is not necessarily a feature that should be avoided. Inclusion of the sample size factor allows for investigation of the "index factor" (i.e., AIC vs RMSEA), in which it is hypothesized that AIC will be sensitive to sample size while RMSEA will largely not be sensitive to it.

3.3.4 Model Complexity

Model complexity will be varied by manipulating the number of indicators for each factor. Data will be generated with either three or six indicators per factor. In either scenario the estimating models will also be adjusted for either three or six indicator per factor. Thus, the low and high model complexity conditions have a total of 11, and 22 observed variables, respectively (both incorporate three latent variables). Past research has

indicated that model fit is affected by the number of indicators per variable (Kenny & McCoach, 2003). RMSEA in particular was shown to improve as the improve as the number of indicators increases.

3.3.5 Measurement Quality

Measurement quality will be varied by manipulating the factor loading indicator strength. Measurement quality has three levels "High", "Medium" and "Low" measurement quality. The average value of the high measurement quality condition is .828. The loading values were inspired by the individual parameter estimates encountered from the Political Democracy example (see Figure. 2.8) and are not uniform. The full list of target parameter values can be viewed in Appendix A2. The average factor loadings for the medium measurement quality was .579, and the average value for the low measurement quality condition was .414. The medium measurement quality condition values are 30% less than that of the high condition while 30% higher than the low condition.

The measurement quality is included as a manipulated factor for two reasons. Structural parameters (i.e., the latent paths) are often the most important parameters to investigators utilizing an SEM framework. Thus, many investigators aim to assess model fit at the measurement level and the structural level separately. If measurement quality is low, it can obscure poor model fit in the structural portion of the model whereas indicators with higher reliability reveal it (McNeish & Hancock, 2018; McNeish, An, & Hancock, 2018). This "reliability paradox" gives an unfortunate incentive (perhaps unknowingly) to prefer unreliable measured variables. It is hypothesized that this relation will be maintained when investigating the FPE range for structural variables. Despite this expectation, it is

unknown to the degree to which this finding will be apparent in the uncertainty measured by FPE analysis.

The total number of proposed model conditions is shown in Table 3.5. These conditions represent the first large simulation study to ascertain the usefulness of the FPE analysis framework from a simulation study rather than previous examinations of real data examples.

Factors	Number of Levels	Value of Levels
Sample Size	3	(200, 1,000, 5,000)
Model Complexity	2	(Low, High)
Model Conditions	3	(Perfect, Misspecified-1, Misspecified-2)
Measurement Quality	3	(.414, .579, .828)
FPE Index	2	(RMSEA, AIC)
Fit decrement	3	RMSEA = (.001, .01) or AIC = 2
Number of Conditions	162	
Total Study Replications	$120 \times 162 = 19{,}440$	

Table 3.5. *Proposed Manipulated Factors.*

This chapter summarizes the methodological approach utilized to investigate FPEs under a variety of conditions. It begins with an overview of the principles used to select conditions, and a presentation of the dissertations research questions. Next this chapter reviews the simulation design, including discussion on the outcome measures, the justification of using a Monte Carlo simulation study, and steps taken to accommodate possible nonconvergence. Finally, this chapter ends with a review of the manipulated factors and their levels included. The following chapter presents the simulation results.

Chapter 4: Results

This chapter presents the studies results in three parts starting with the highest level of summary and ending with the most granular. The overall results section provides the highest level of summary containing only a single outcome per condition – the mean study outcome across all parameters. Next, the results by parameter type subgroup are presented in which the mean outcome is displayed by parameter type. Then the Results by Individual Parameter section (4.2.4) presents results at the most granular level. FPE analysis is, however, likely best interpreted at the individual parameter level. Because of the sizable quantity of results generated and the complexity of the study it is useful to first begin with more summative results. These two introductory sections will also additionally highlight factors that were ostensibly important at the higher levels of summary so that they might be investigated further at the parameter level of analysis.

While the majority of the interpretation of the results and their meaning in reference to past and future studies is found in the discussion chapter, this chapter ends with a summative overview of the results by manipulated factor. Finally, preceding the primary results is a data evaluation section that reports the level of precision in the study outcome.

4.1 Data Evaluation

4.1.1 MCE ML Estimates

The data were evaluated to determine whether there was a sufficient level of precision across the conditions. This was done by calculating the MCE of the estimates (Equation 3.3). The MCE is calculated for each parameter estimate in the model (for a total ranging between 35 and 64 depending on the model complexity condition which was largely defined as the number of model parameters). The mean of these individual MCE estimates was then calculated and is reported as a single summary for each of the 102 RMSEA conditions in Figure 4.1. Here MCE for each MLE is displayed. Although the MLEs are not a direct study outcome variable, it is important to determine that they are simulated precisely for two reasons. First, it ensures that the study conditions were indeed simulated correctly. Second, both metrics of parameter stability – the FPE percent change and the FPE range metric – require that the MLE are measured accurately. The mean estimated MCE for the ML parameter estimates across all conditions is 0.004, and the maximum is 0.019.

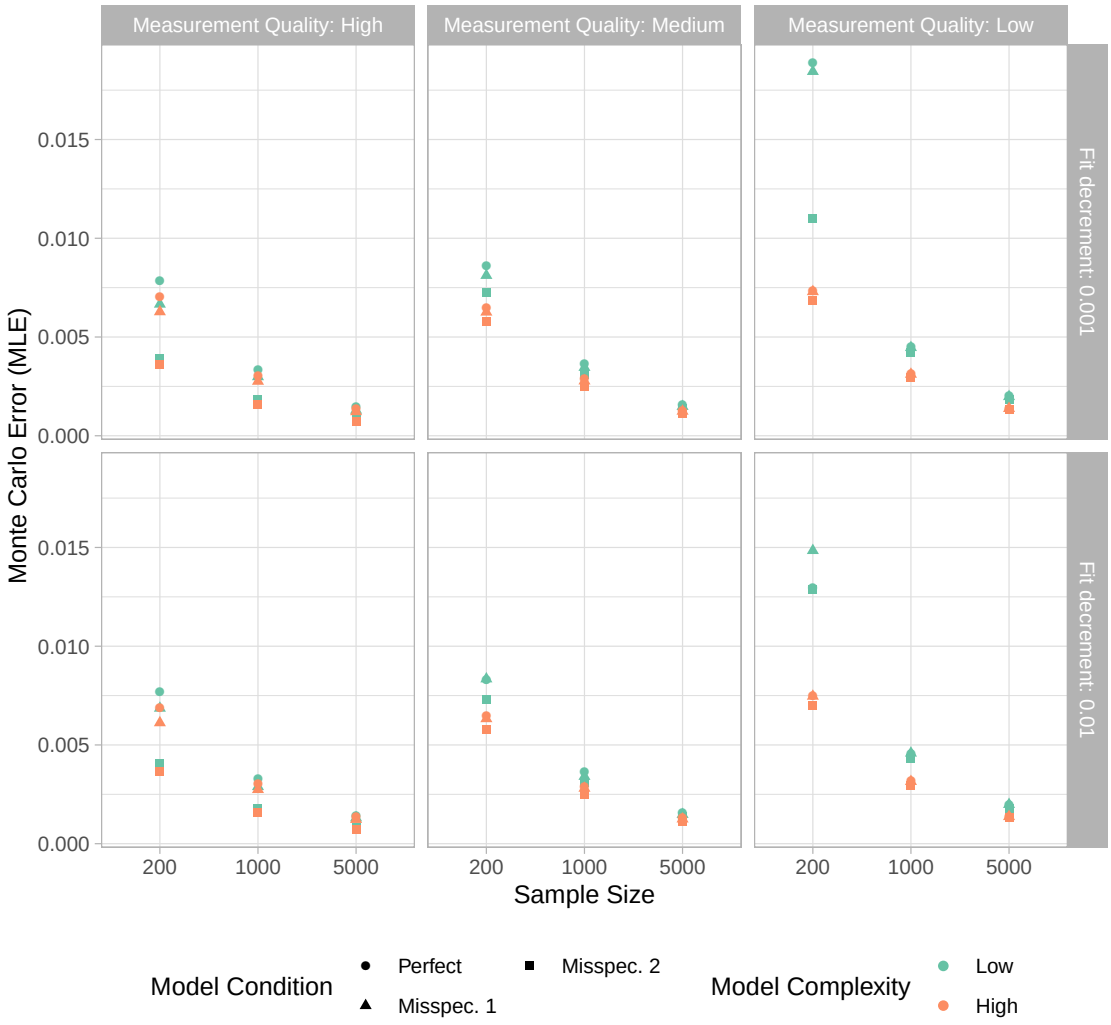

Figure 4.1. Estimated MCE for ML Parameter Estimates. Each point represents the mean MCE across all parameter estimates for that respective condition.

4.1.2 MCE FPE Percent Change Metric

The precision of the FPE percent change metric was also calculated. Valid interpretation of the percent change metric rests on the notion that the values presented in the results section do represent the population values (i.e., they must be unbiased and measured with sufficient precision). In order to examine this, the MCE was once again calculated, as described in section 3.2.4.2 . The mean estimated MCE value for each of the 102 RMSEA conditions (i.e., those that use RMSEA as a measure of model fit) is shown in

Figure 4.2. The mean estimated MCE across all conditions was 62.2, and the maximum was a very high 1927 (see Figure 4.2).

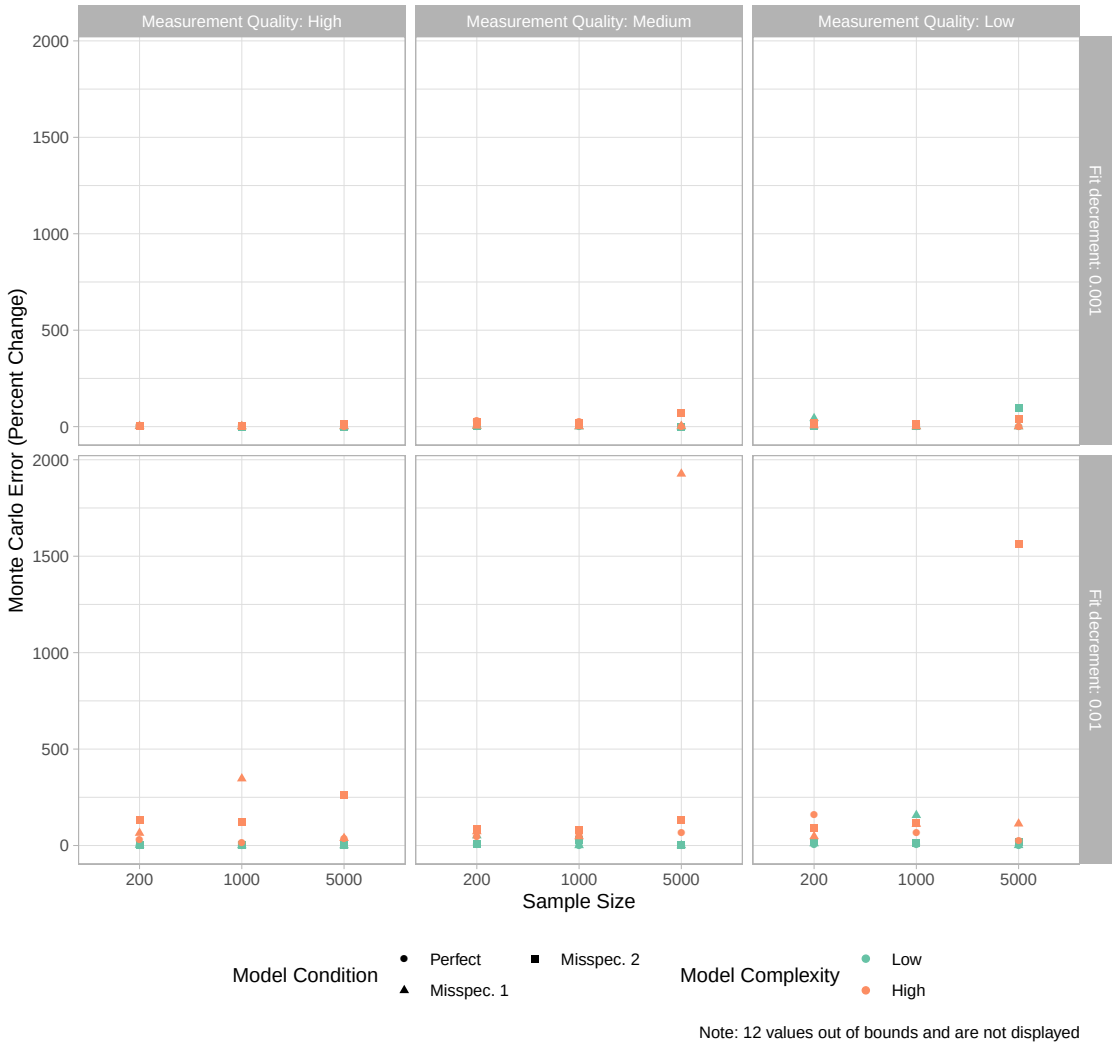

Figure 4.2. Estimated MCE for FPE percent change metric. Each point represents the mean parameter estimate MCE for a single condition.

4.1.3 MCE FPE Range Metric

Finally, the MCE for the raw FPE range was also calculated. The results of this analysis, by condition, are shown in Figure 4.3. The mean estimated MCE across all conditions is 0.005, and the maximum is 0.022.

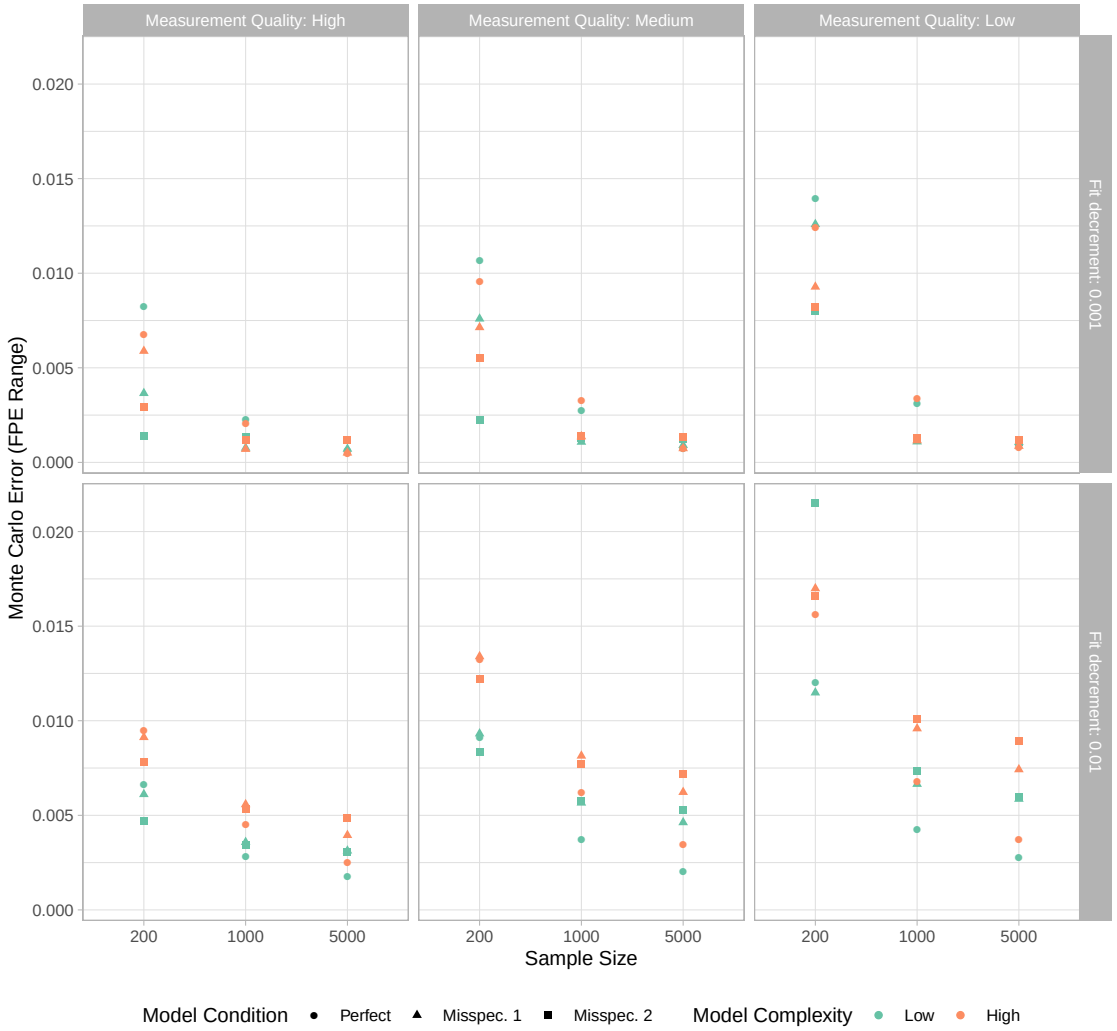

Figure 4.3. Estimated MCE for the range of fungible parameter estimates. Each point represents the mean parameter MCE for a single condition.

4.2 Outcome by Level of Summary

4.2.1 Overall Results

This section provides a summary of the factors on the FPEs when averaged across all estimated model parameters.

4.2.1.1 Exploring Results with Eta-Squared. A factorial ANOVA was used to apportion the variability in the outcome of the simulated data. Specifically, $\hat{\eta}^2$ is reported to quantify the effects of the different study factors. The proportion of total variance

accounted for by a particular factor (or interaction of factors) is represented by $\hat{\eta}^2$. This approach is meant as a tool in order to both discover potentially important interactions (and filter out others) so that they can be analyzed in the results by individual parameter section (4.2.4).

The initial model contained all interaction terms. Because of the difficulty in interpreting higher level interactions this model was substituted for a less complex one. The largest $\hat{\eta}^2$ for a three-way interaction (or greater) was 0.004. The six-four way interactions each described less than .001 percent of the variance in the model output. In addition, four-way interactions would likely have limited utility for researchers as they are often difficult to interpret and thus will also not be analyzed further. Further values of this initial analysis can be found in Appendix B (see Tables B5, B6, and B7).

The final models included manipulated study factor main effects and two way interactions. Higher order interactions were included in the error term. Details of the variables included in the model can be found in the respective table notes. The results of the final model are shown in Tables 4.1, 4.2, and 4.3 and are discussed in upcoming sections.

4.2.1.2 RMSEA and AIC. Table 4.1 displays $\hat{\eta}^2$ for all factor effects and interactions (including up to two way interactions). The total proportion of variance accounted for by factors was 0.319. Figure 4.4 displays all factors—sans the model complexity condition. This condition was omitted for figure clarity, and because it described the least amount of variability of any single model factor (i.e., 0.005). Overall, the largest ranges were found in the RMSEA = .01 condition. The sample size condition has a differential effect on FPE ranges depending on the index of model decrement used.

Table 4.1. $\hat{\eta}^2$ *for Factors Representing the Manipulated Conditions*

	Factor	$\hat{\eta}^2$
1	Sample Size	0.04
2	Model Complexity	0.01
3	Model Condition	0.05
4	Measurement Quality	0.01
5	Fit Index	0.16
6	Sample Size : Model Complexity	0.00
7	Sample Size : Model Condition	0.01
8	Sample Size : Measurement Quality	0.00
9	Sample Size : Fit Index	0.00
10	Model Complexity : Model Condition	0.00
11	Model Complexity : Measurement Quality	0.00
12	Model Complexity : Fit Index	0.01
13	Model Condition : Measurement Quality	0.00
14	Model Condition : Fit Index	0.02
15	Measurement Quality : Fit Index	0.00

$\hat{\eta}^2$ was calculated for two-way interactions and main effects for the following factors: Sample Size, Model Complexity, Model Condition, Measurement Quality, and Fit Index. All higher order interactions were included in the error term. Note: Model Decrement is not included.

Also from this figure, the relation between model misspecification condition is different between levels of model fit decrement (i.e., RMSEA = 0.01 vs RMSEA = 0.001). For the RMSEA = 0.001 level the "Perfect" condition features the greatest FPE ranges, whereas the "Misspec. 1" and "Misspec. 2" conditions have the greatest FPE ranges for the RMSEA = 0.01 level. Finally, on average the "Low", and "High" Measurement Conditions have the largest and smallest FPE ranges, respectively.

As argued previously, the two "fit indexes" of FPE decrement represent two different methods of measuring uncertainty in an FPE analysis. These two different indexes are conceptualized to represent differing ways of measuring uncertainty, rather than being thought of as a traditionally manipulated factor level. For this reason, the next analysis

subset the results by AIC and RMSEA (rather than as combined in Table 4.1).

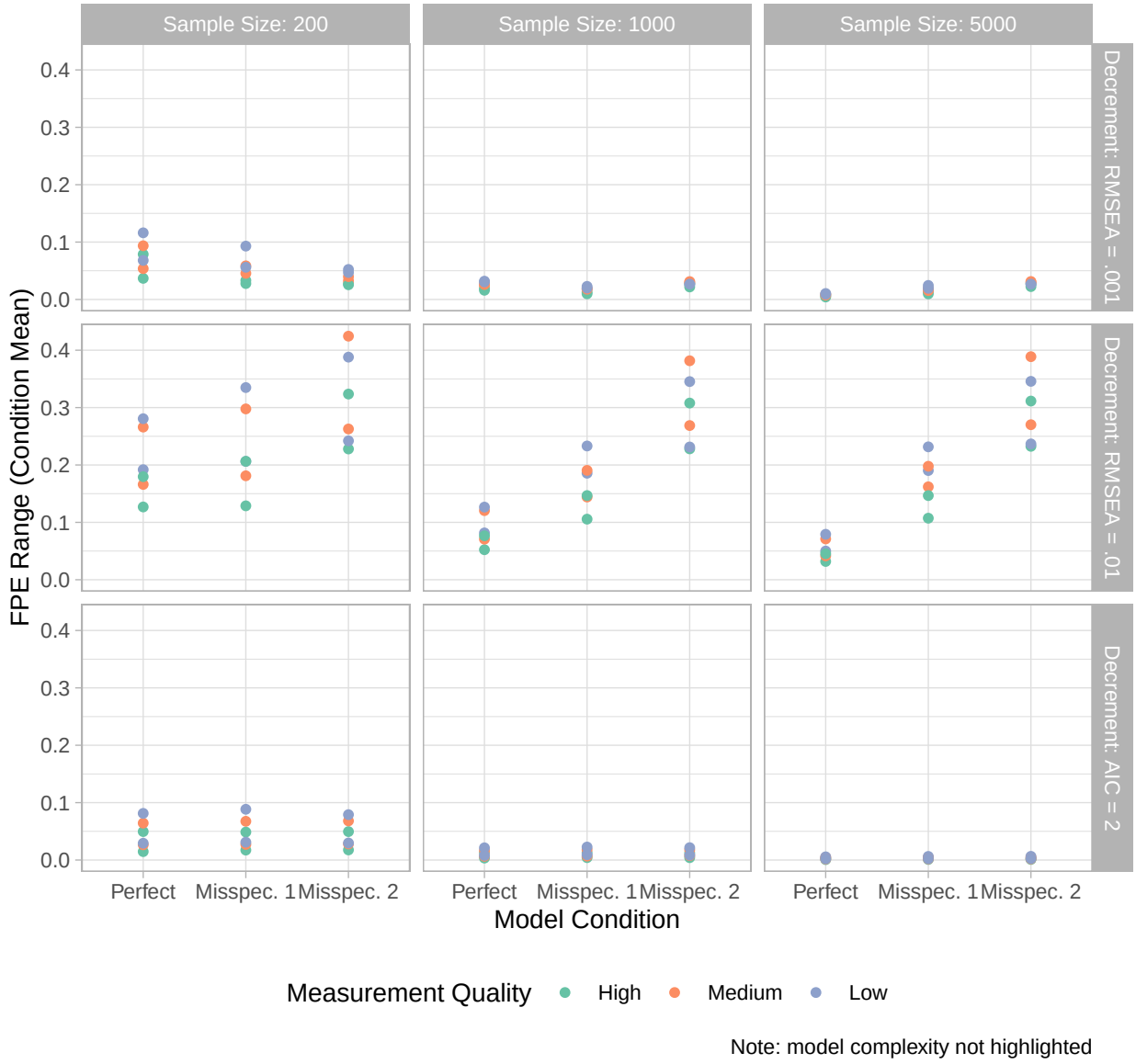

Figure 4.4. FPE range by study factor. Each point represents the mean FPE range averaged across study condition and all model parameter.

4.2.1.3 RMSEA. This section includes only those conditions that use RMSEA as a measure of model fit decrement (N = 108). Table 4.2 shows the proportion of total variability in range of FPE estimates by factor. The total proportion of variance accounted for by all factors utilizing RMSEA as a measure of model fit decrement is 0.776. For

RMSEA, the fit decrement condition explains more than the other factors combined. This indicates that the larger the decrement the larger the fungible parameter range expected. Interestingly, it can be seen that fit decrement has interactions with several other conditions. The interactions with model condition, and model complexity indicate that the relation between the FPE ranges and these variables can differ depending on the size of the fungible parameter decrement selected. This finding is analyzed further in the Results by Individual Parameter section (4.2.4).

Table 4.2. $\hat{\eta}^2$ *Factors Representing the Manipulated Conditions (RMSEA)*

	Factor	$\hat{\eta}^2$
1	Sample Size	0.04
2	Model Complexity	0.02
3	Model Condition	0.09
4	Measurement Quality	0.02
5	Fit Decrement	0.45
6	Sample Size : Model Complexity	0.00
7	Sample Size : Model Condition	0.02
8	Sample Size : Measurement Quality	0.00
9	Sample Size : Fit Decrement	0.00
10	Model Complexity : Model Condition	0.00
11	Model Complexity : Measurement Quality	0.00
12	Model Complexity : Fit Decrement	0.03
13	Model Condition : Measurement Quality	0.00
14	Model Condition : Fit Decrement	0.10
15	Measurement Quality : Fit Decrement	0.01

$\hat{\eta}^2$ was calculated for two-way interactions and main effects for the following factors: Sample Size, Model Complexity, Model Condition, Measurement Quality, and Fit Decrement. All higher order interactions were included in the error term. Note: Fit Index is not included.

Figure 4.5 displays two prominent interactions with Fit Decrement. The Model Complexity*Fit Decrement and Model Condition*Fit Decrement interactions. For the former, the high complexity conditions have smaller FPE ranges for the RMSEA = .001 condition, while the low complexity conditions have the smaller FPE ranges for the RMSEA = .01 conditions. The latter interaction—between Model Condition and Fit Decrement–demonstrates that the size of FPE ranges are not necessarily predictable when comparing several different models at only one level of model fit decrement. This interaction is most clearly demonstrated for the RMSEA = .001, n = 200 condition. The RMSEA = .001 conditions shows larger FPE ranges for the perfect conditions and, in several cases, smaller ranges for the more poorly fitting misspecified conditions. However, the perfect model demonstrates the smallest FPE ranges for the RMSEA = .01 condition.

4.2.1.4 AIC. The following section features the results for the 54 conditions that used AIC as a measure of model fit decrement. As shown in Table 4.3, the total amount of variability in these conditions explained by all factors is 0.641. For AIC, 62% of this explained variance is from the sample size factor. Figure 4.6 displays the direction of these effects. In all cases examined, the high complexity conditions have smaller FPE ranges compared with that of the low complexity conditions. The Sample Size * Model Complexity interaction is instantiated as large increases in stability for the small sample size, but smaller decreases for the larger sample size. Conditions with higher measurement quality have smaller FPE ranges than medium or low measurement quality conditions.

4.2.2 Results by Parameter Type Subgroup

The results for the previous section represent an average of all parameter estimates

Table 4.3. $\hat{\eta}^2$ *Factors Representing the Manipulated Conditions (AIC)*

	term	etasq
1	Sample Size	0.40
2	Model Complexity	0.10
3	Model Condition	0.00
4	Measurement Quality	0.02
5	Sample Size : Model Complexity	0.09
6	Sample Size : Model Condition	0.00
7	Sample Size : Measurement Quality	0.02
8	Model Complexity : Model Condition	0.00
9	Model Complexity : Measurement Quality	0.00
10	Model Condition : Measurement Quality	0.00

$\hat{\eta}^2$ was calculated for two-way interactions and main effects for the following factors: Sample Size, Model Complexity, Model Condition, and Measurement Quality. All higher order interactions were included in the error term. Note: Model Index, and Fit Decrement were not included.

within a study condition. The next sections examines the relation of the study factors for subsets of model parameters to investigate whether the study factors deferentially affect these subsets of parameters.

This analysis begins with Figures 4.7 and 4.8 which like previous figures subset the results by FPE index. Figure 4.7 displays the RMSEA indexed results. For the RMSEA conditions, overall the (co)variance parameters are the most stable. For the RMSEA = 0.01 condition, (co)variance parameters are the most stable for the perfect condition. However, this pattern is reversed for many of the (co)variance parameters in the RMSEA = .001 fit condition. This reflects a similar interaction between model fit decrement and model condition shown previously in the overall results section (see 4.2.1 and Figure 4.5). This interaction is not expected and the degree to which it holds true at the parameter level is shown in the results by individual parameter section (4.2.4).

Overall, both Figure 4.7 and 4.8 display the breakdown of FPE ranges by type of

variable and show the same pattern as when this dimension was collapsed in the overall results section (Figures 4.5 and 4.6). Subsetting the results by parameter type did, however, indicate that the (co)variance parameters are the most stable for both RMSEA and AIC conditions.

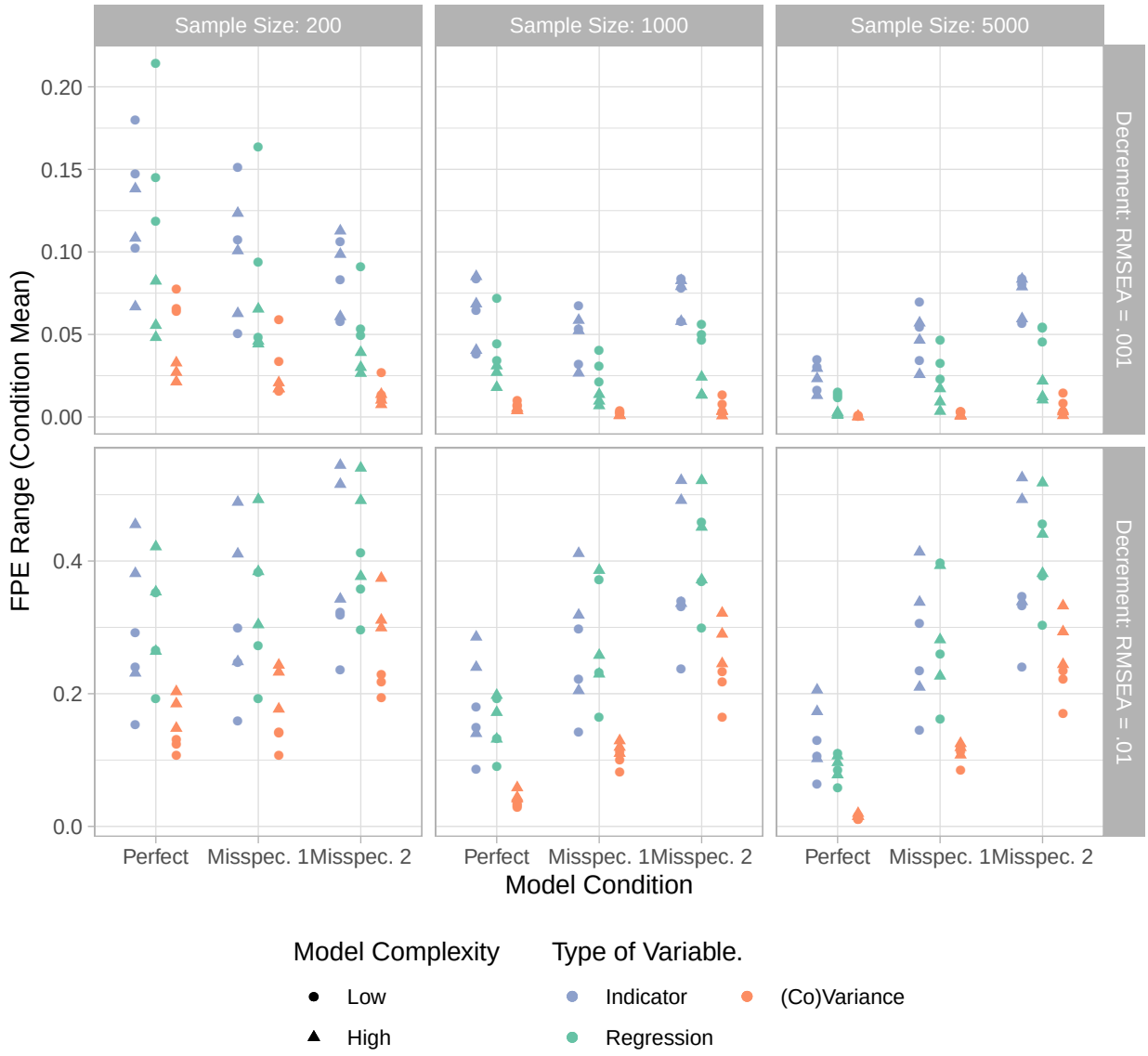

Figure 4.7. FPE range by variable type for conditions that used RMSEA as measure of model fit. Each point represents the mean FPE range averaged across study condition

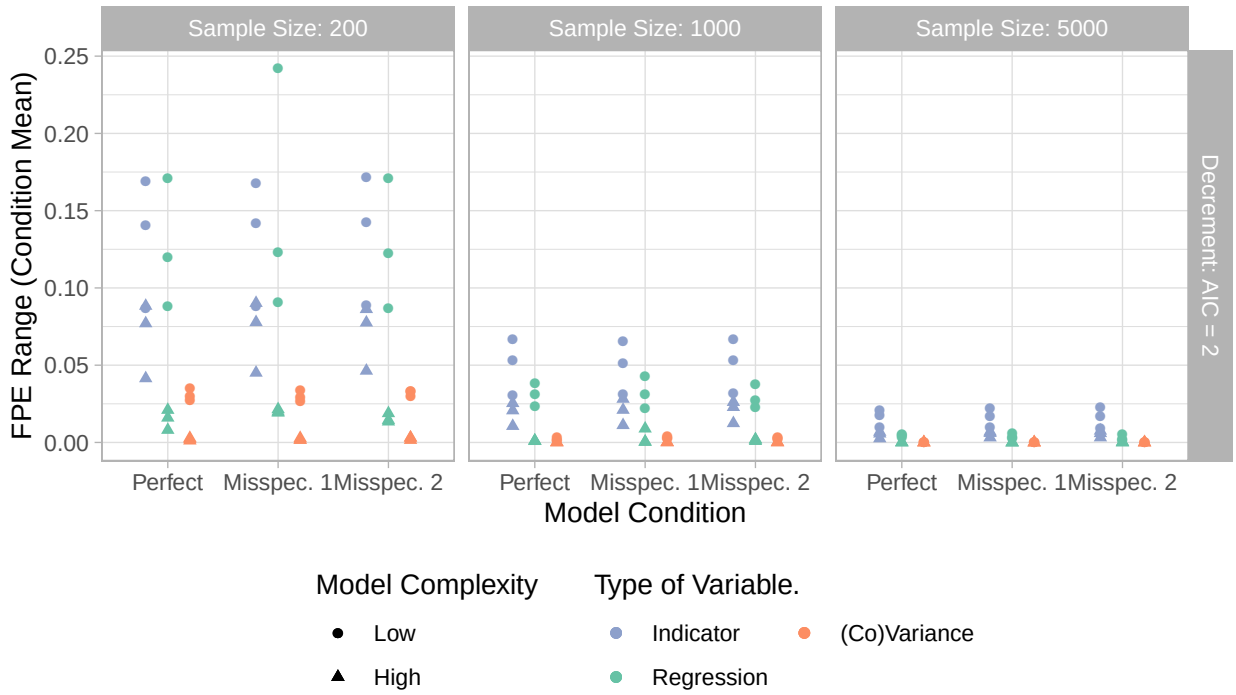

Figure 4.8. FPE range by variable type for conditions that used AIC as measure of model fit. Each point represents the mean FPE range averaged across study condition

4.2.3 Summary of Overall and Parameter Type Subgroup Results

In the overall results section, there is a large $\hat{\eta}^2$ for the effect of Fit Index (i.e., the difference between AIC and RMSEA). Some of this difference is due to the RMSEA decrement condition containing the .01 condition collapsed with the .001 condition. The RMSEA = .01 condition is often times much worse fitting than the AIC = 2 condition as measured by raw objective function values. Because of this, it was useful to conduct $\hat{\eta}^2$ analysis by the index of model decrement separately. It was a goal to understand which factors are most responsible for FPE ranges at more summative levels of analysis. However, these analysis were also aimed at uncovering interactions between factors which might be investigated at the individual parameter level. In the overall results section, the Model Condition*Fit Decrement and Model Complexity*Fit Decrement were important interactions for RMSEA indexed FPEs. Sample size and the Sample Size*Model Complexity factors were important for AIC indexed conditions when using the average size of all FPEs across parameters.

A similar set of patterns is also apparent when the results are subset by parameter type. For example, the interaction Model Condition*Fit Decrement for RMSEA. These figures do, however, reveal another interaction between parameter type and Fit Decrement. Covariance parameters for the RMSEA = .001 condition are the most stable, whereas they are often the least stable for the RMSEA = .01 condition.

The next section examines the impact of the factors on FPE ranges at the level of individual parameters. The previous two sections suggest a usefulness in subsetting results by FPE Index. In addition, the Model Condition*Fit Decrement, Model Complexity*Fit Decrement, and Model Fit Decrement*Parameter Type factors for RMSEA were important. While Sample Size*Model Complexity interaction was important specifically for

AIC indexed FPEs. These four interactions now are joined with the main effects shown in Table 3.5 (i.e., Sample Size, Model Complexity, Model Misspecifications, Measurement Quality, FPE Index, and Fit decrement) for further consideration.

4.2.4 Results by Individual Parameter

The preceding sections displayed results that were generated by calculating the mean values across all model parameters, and subset by parameter type subgroup (i.e., indicator, regression, or (co)variance parameters), respectively. This information was meant to give an overview of the results and help to demonstrate the main effect of each study factor at a top level. However, in a generalized FPE analysis parameter uncertainty and sensitivity can be quantified at the individual parameter level. Due to the differing number of parameter estimates between high and low complexity conditions some parameters will necessarily have differing numbers of estimates. For example, the "dem65_y22" parameter occurs in half of the 162 conditions whereas "dem65_y11" exists in every condition. The upcoming section focuses on Sample Size.

4.2.4.1 Sample Size. The results shown in Figure 4.9 that increasing sample size decreases FPE ranges. The effects are not uniform across parameters, however. There is a different order as to which parameters have the largest instability for RMSEA and AIC indexed FPEs, but the overall pattern is similar. For both indexes, (co)variance parameters have smaller FPE ranges, while indicator, and latent regression parameters have larger FPE ranges.

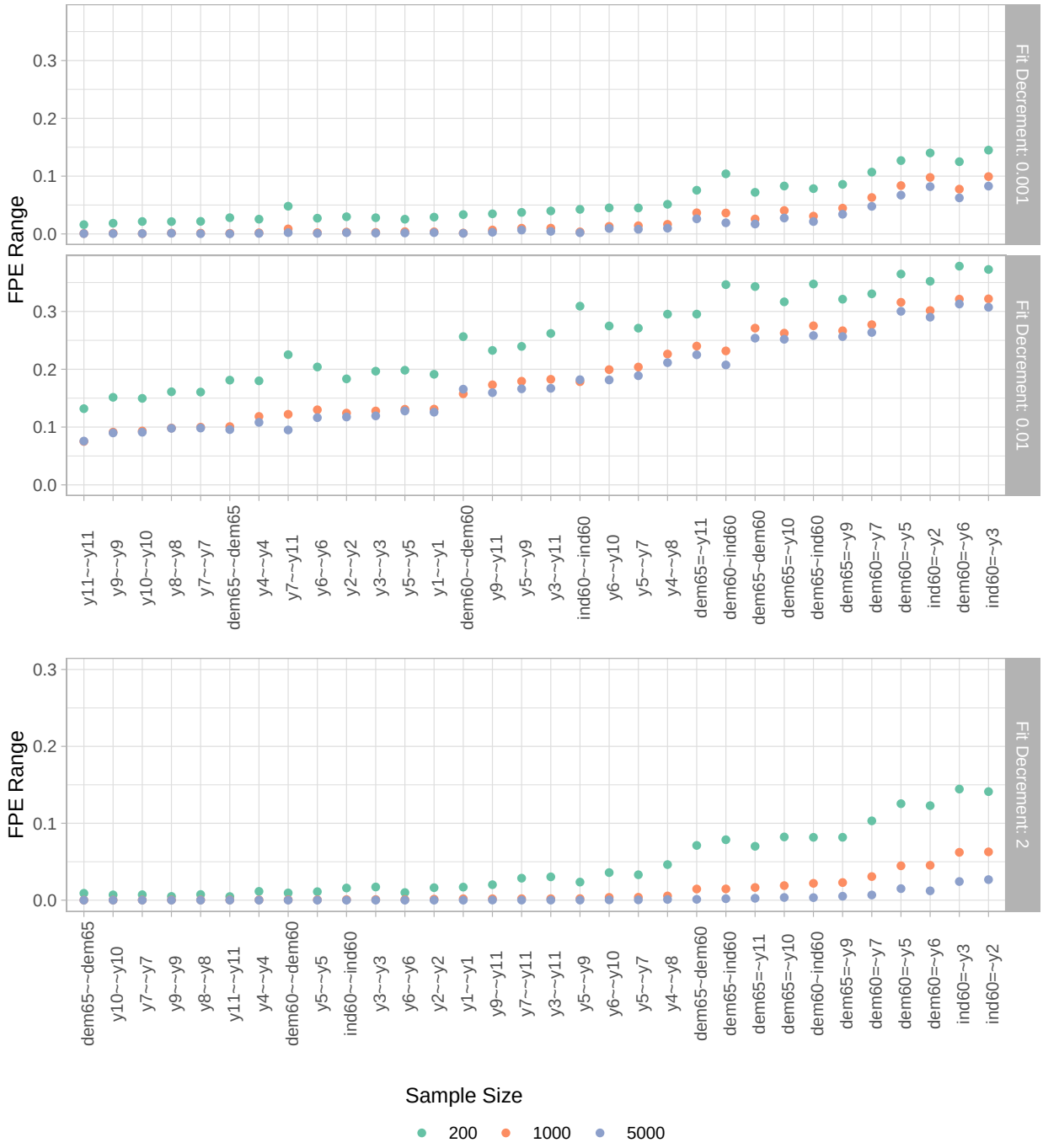

Figure 4.9. FPE range by parameter for both AIC and RMSEA indexed conditions. Each parameter has three different points indicating the stability of that parameter at n = 200, 1000, and 5000 sample sizes at the given level of fit decrement

4.2.4.2 Model Complexity.

For this study, model complexity is defined by differing number of indicators for each latent variable. The low complexity condition

contains 11 observed variables, whereas, the high complexity condition contains 22 observed variables. The overall effect of model complexity when collapsed across model index (i.e., RMSEA vs. AIC) is that the low model complexity conditions are more stable for the majority of parameters (see Appendix B1). Secondly, this pattern is relatively consistent across model conditions. However, when subset by model index, the results were split. For RMSEA indexed conditions the FPE ranges are more stable, but not for AIC-indexed conditions (Figure 4.10).

Figure 4.10. FPE range by parameter for both AIC and RMSEA. The bars are colored blue when the low complexity condition is relatively more stable than the high complexity condition, and red when the high complexity condition is more stable than the low complexity condition. Only parameters that are included in both the high and low complexity conditions are included. Note the differing y-axis scales.

4.2.4.3 Model Complexity*Fit Decrement (RMSEA).

The effect of model complexity as shown previously (see Figure 4.5), is not uniform across model fit decrement. Figure 4.11 allows for a determination of whether this interaction is consistent across

parameters. The results suggest that this pattern is indeed largely consistent. For AIC-indexed conditions, every parameter is more stable for the high complexity condition. For the RMSEA conditions there are a few exceptions. For the smaller decrement, RMSEA = .001, the majority of parameters are more stable for the high complexity conditions. Of these parameters 13/93 indicated that the low complexity condition is more stable. All 13 of these parameters represent latent variable indicators. For the RMSEA = .01 conditions, most parameters indicated that the low complexity condition was more stable. For this condition, only 4/93 parameters are exceptions to the pattern (more stability for the high complexity condition). Unlike the smaller decrement condition, RMSEA = .001, these exceptions to the overall pattern represent some of the smallest differences between the small and complex modeling conditions (i.e., the 1st and 4th smallest changes in model stability).

Figure 4.11. FPE range by parameter for both AIC and RMSEA. The high and low complexity conditions for AIC (top row). The bars are colored blue when the low complexity condition is relatively more stable than the high complexity condition, and red when the high complexity condition is more stable than the low complexity condition. Only parameters that are included in both the high and low complexity conditions are included

4.2.4.4 Model Complexity*Sample Size (AIC).

Increasing sample size decreases the fungible parameter range of estimates for all parameters. For AIC indexed conditions, the sample size effect interacts with model complexity. This interaction was less salient in previous sections (see 4.6 [overall results] and 4.8 [by parameter type]). Figure 4.12 indicates that FPE ranges are reduced for the high complexity condition, whereas the differences (i.e., between model complexity conditions) for indicator variables are relatively large as sample size increases. The three latent parameters (i.e., dem65 dem60, dem65 ind60, and dem60 ind60) also maintain relatively large differences between complexity conditions at the high sample size when compared to variance parameters. That is, as sample size increases only a subset of parameters maintain relatively unstable ranges for the low complexity condition.

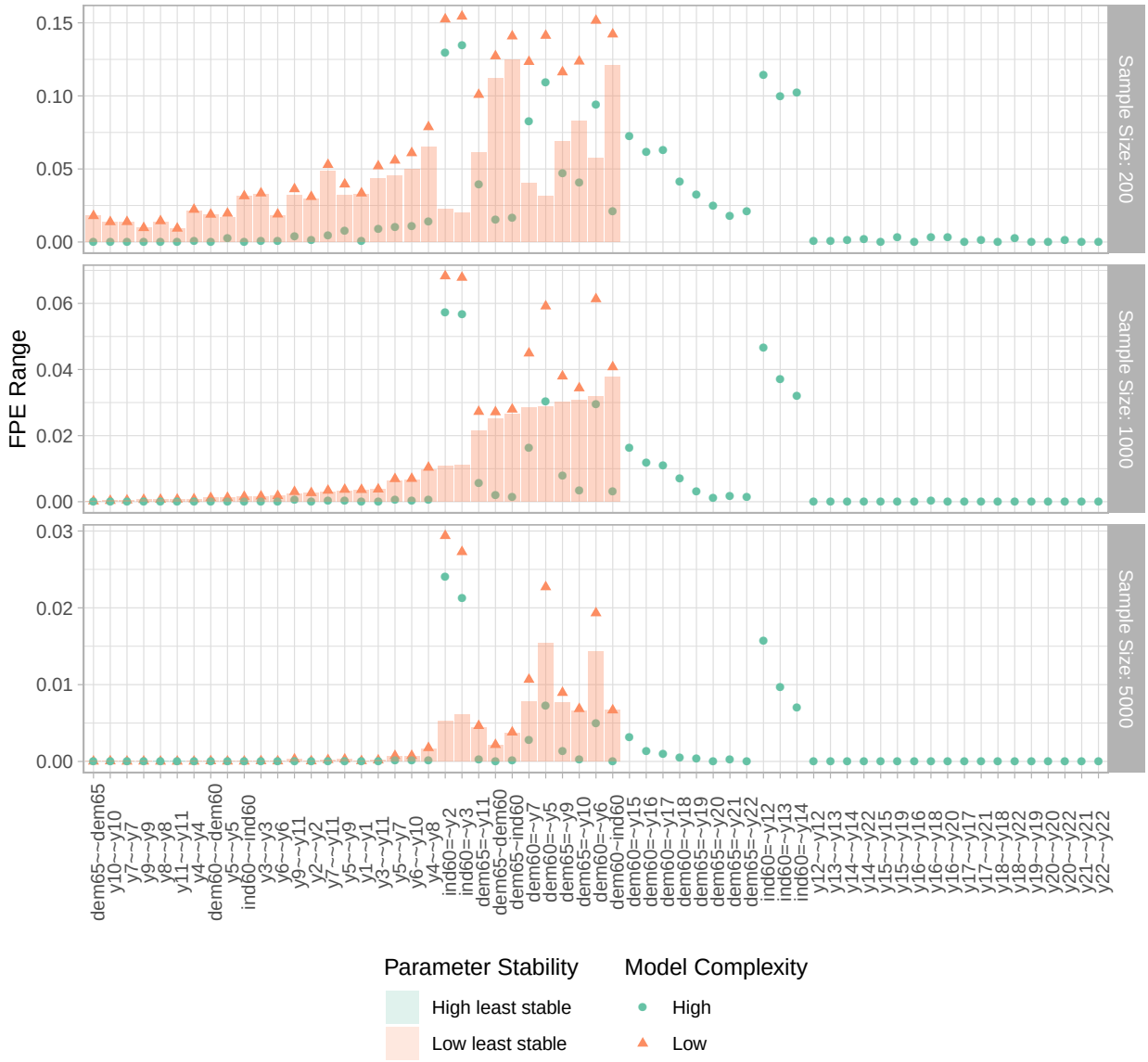

Figure 4.12. FPE range by parameter for AIC. Parameters have up to two different points representing the stability of that parameter at the high or low model complexity conditions. The bar heights represent the magnitude of the difference between these two conditions. The parameters without a second set of points only appear in the high model complexity conditions. Note the differing y-axis in each facet.

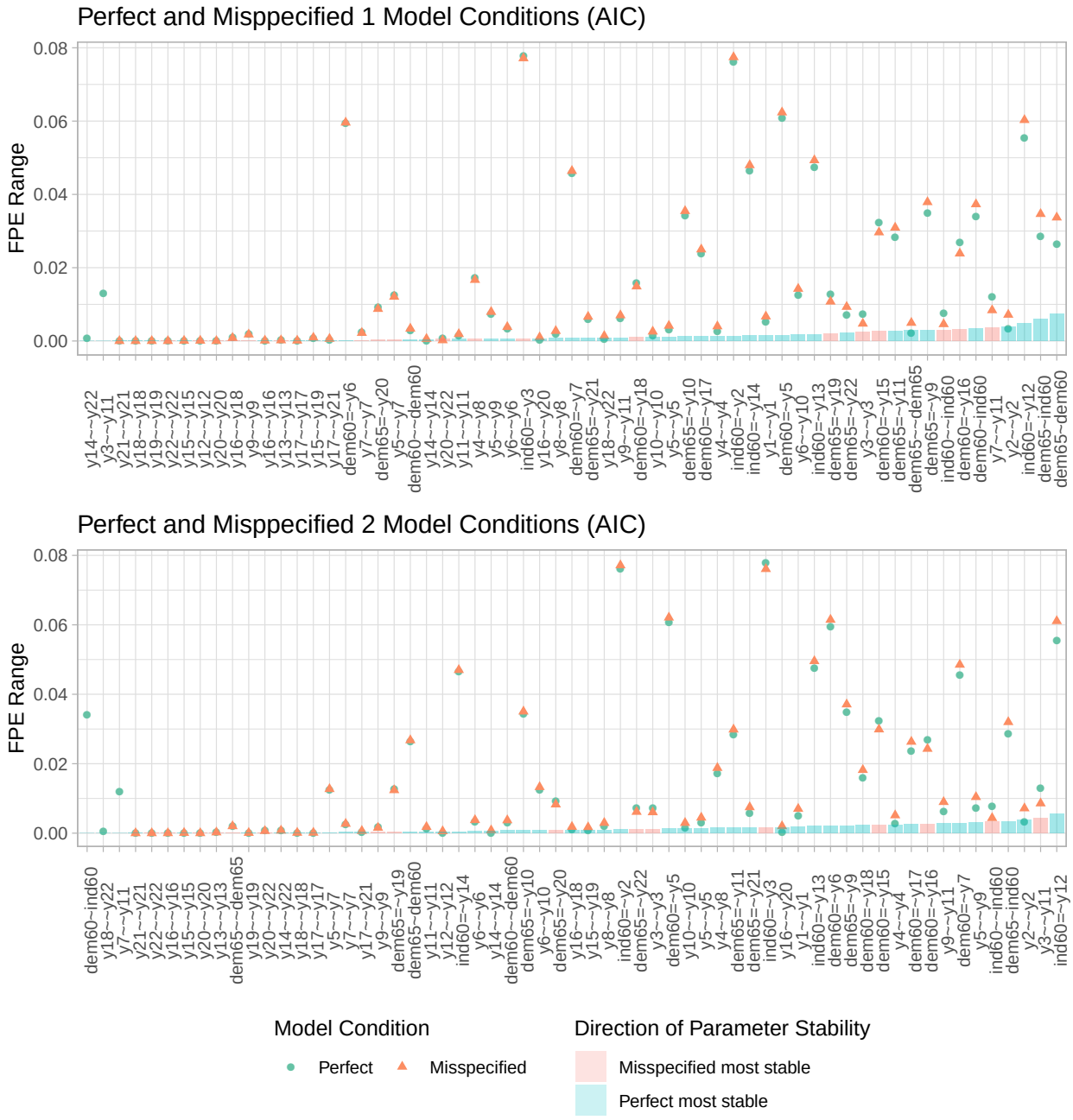

Figure 4.13. FPE range by parameter for AIC. Variables are ordered by the size of the difference in parameter stability between the perfect condition and the misspecified 1 (top facet) and misspecified 2 (bottom facet) conditions. The bars are colored blue when the perfect condition is relatively more stable than the respective misspecified condition, and red when the misspecified condition is more stable than the perfect condition.

4.2.4.5 Model Condition.

Figure 4.13 demonstrates the main effect of model misspecification at each parameter for AIC indexed conditions. Because there is only one

level of model decrement for AIC, both misspecification conditions are included in a single figure. For AIC, the majority of parameters are more stable for the perfect condition. This pattern is similar to the larger of the two RMSEA fit decrement conditions (see Figures 4.15 and 4.16).

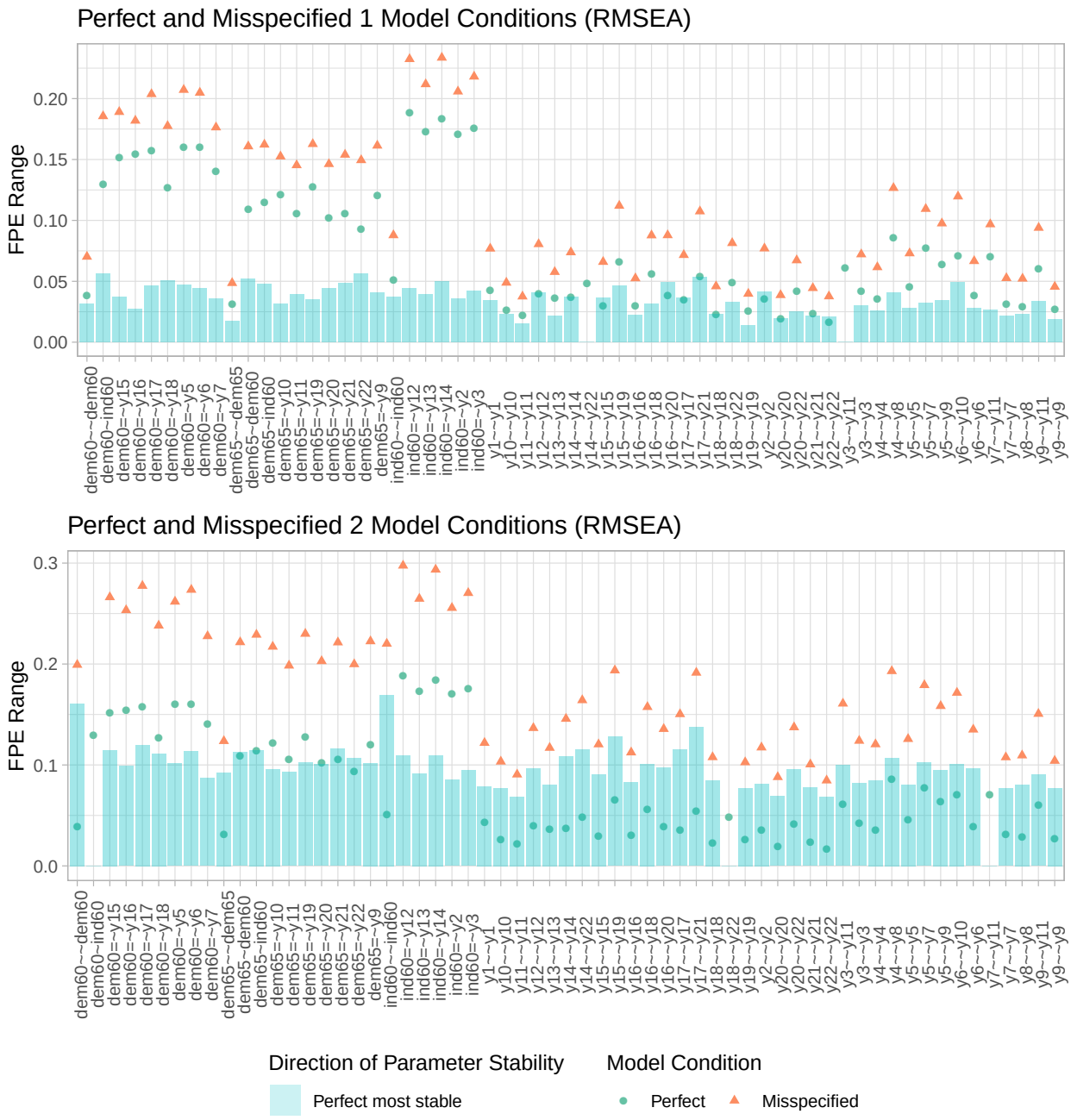

Figure 4.14. FPE range by parameter: perfect and misspecified 1 (top facet) and 2 (lower facet) conditions for RMSEA. The bars are colored blue when the perfect condition is relatively more stable than the respective misspecified condition, and red when the misspecified condition is more stable than the perfect condition (not existent here).

Figure 4.14 demonstrates a similar plot for RMSEA indexed conditions. When compared to either misspecification conditions, the perfect condition is the most stable

model condition for all parameters. This result is in line with the expectation that the perfect condition would serve as the most stable of the three conditions. This consistent pattern of results is contrasted with the upcoming RMSEA indexed results which demonstrate a diverging pattern based on the level of fit decrement.

Figure 4.15. FPE range by parameter: perfect and misspecified 1 conditions for both decrement levels for RMSEA. Variables are ordered by the size of the difference in parameter stability between perfect and misspecified 1 conditions (i.e., length of the bars). The bars are colored blue when the Perfect condition is relatively more stable than the misspecified 1 condition, and red when the misspecified 1 condition is more stable than the perfect condition.

Figures 4.15 and 4.16 display the interaction between decrement size and model misspecification. This section only applies to RMSEA indexed conditions because there is only one level of fit decrement for AIC indexed conditions. At the larger decrement size

(i.e., RMSEA = .01) the perfect condition is more stable but not at the smaller decrement size (i.e, RMSEA = .001).

4.2.4.6 Model Condition*Fit Decrement (RMSEA). While these results are consistent with the overall results presented earlier (see 4.5 and 4.7), the benefit of figures 4.15 and 4.16 is that they enable examination of whether these effects are uniform across parameter estimates. There is a consistent pattern of results for the 0.01 fit decrement level in which every parameter is more stable for the perfect model condition. However, there are a number of exceptions at the 0.001 fit decrement level in that not every misspecified parameter is more stable. For the misspecification 1 condition (Fig. 4.15) there are three exceptions in which the perfect conditions are more stable. For the misspecification 2 condition (Figure 4.16) the results are more mixed with 15/58 variables indicating more stability for the perfect condition.

Figure 4.16. FPE range by parameter: perfect and misspecified 2 conditions for both decrement levels for RMSEA. Variables are ordered by the size of the difference in parameter stability between perfect and misspecified 2 conditions (i.e., length of the bars). The bar are in colored blue when the perfect condition is relatively more stable than the misspecified 2 condition, and red when the misspecified 2 condition is more stable than the perfect condition.

4.2.4.7 Measurement Quality.

The effect of measurement quality (the reliability of indicator variables) appears in Figures 4.17 and 4.18. These two figures compare FPE ranges of misspecified conditions 1 and 2 against the perfect condition, respectively. These

figures also display each level of model decrement.

The model misspecification 1 condition misspecifies a null relationship between y3 and y11 for which the effect is present in Figure 4.17. In Figure 35, the more poorly fitting RMSEA = .01 condition differs from the other model decrement conditions in that every parameter becomes more unstable irrespective of the measurement quality condition. A further study of the misspecification conditions follows with a focus on a subset of parameters of interest to guide the analysis.

4.2.4.8 Effect of Model Misspecification, Measurement Quality, and FPE Index on Parameters of Interest.

Misspecification 1. The misspecified 1 condition is misspecified because it is missing the path between y3 and y11. This true covariance between y3 and y11, however, is still present in the data generation process. Because of this, other sections of the model that are estimated must take into account this misspecified covariance. Understanding how misfit manifests itself in other parts of the model is difficult to predict even in relatively simple models. One method of predicting how this misspecification might manifest itself is by decomposing the model using the rules for latent variable path tracing. The covariance between y3 and y11 can be parsed into three components shown below. The direct path 1a, and the two indirect paths 1b, and 1c. (the summation of these three paths is equal to the model implied covariance of y11_y3).

1a. $y3\_y11$ (ψ_{3_11})

1b. $ind60\_y3 * ind60\_ind60 * dem65\_ind60 * dem65\_y11$ (i.e., $\lambda_{31} * \psi_{11} * \beta_{31} * \lambda_{11_3}$)

1c. $ind60\_y3 * ind60\_ind60 * dem60\_ind60 * dem65\_dem60 * dem65\_y11$ (i.e., $\lambda_{31} * \psi_{11} * \beta_{21} * \beta_{32} * \lambda_{1_3}$).

These three paths represent the covariance of the true model. In the misspecified 1 condition direct path 1a is not specified. The indirect paths, are however, still included in the misspecified 1 model and represent candidates by in which the covariance of $y3\_y11$ may manifest itself if the direct path is not estimated. The following section refers to the results and three levels of fit decrement shown in Figure 4.17.

Figure 4.17 displays the FPE ranges for the misspecified 1 condition with the perfect condition serving as a baseline. These results are subset by level of fit decrement and are described below.

For AIC = 2 conditions the dem60_ind60, dem65_dem60, and dem65_ind60 parameters demonstrate increased FPE ranges, whereas the ind60_ind60 and ind60_y3 decreased in general compared to the perfect condition. The increased FPE ranges for the two latent paths dem65_dem60, and dem65_ind60 represent the largest changes across all model parameters. For these same two latent variables, this increase is largely driven by the low measurement quality condition. Conversely, the low measurement quality covariance ind60_ind60 demonstrated one of the largest increases in stability.

For RMSEA = .001 conditions the majority of the parameters for the misspecified 1 condition become more stable when compared to the perfect condition. While there are differences at the parameter level by measurement quality the pattern is not easily discernible.

For RMSEA = .01 conditions all parameters demonstrate increased FPE ranges across all measurement quality conditions for the misspecified 1 condition compared to the perfect condition. Among the focus variables, the FPE ranges are largest for the low measurement quality conditions.

Misspecification 2. The misspecification 2 conditions contain two misspecifications. The first, is the ommited covariance y11_y7. These two variables co-vary based on a direct path and three indirect paths:

2a. $y11\_y7$

2b. $dem60\_y7 * ind60\_dem60 * ind60\_ind60 * ind60\_dem65 * dem65\_y11$

2c. $dem60\_y7 * ind60\_dem60 * ind60\_ind60 *$
$ind60\_dem60 * dem60\_dem65 * dem65\_y11$

2d. $dem60\_y7 * dem60\_dem60 * dem60\_dem65 * dem65\_y11$

The second misspecification for this model is the path between ind60_dem60. This variable along with ind60_dem65, and y3_y11 are the three variables that associate the measured variables y1, y2, and y3 with measured variables y4-y11 (for the 11 variable non-complex model). The remaining two variables linking the observed variables, ind60_dem65, and y3_y11 will be analyzed along with the variables contained in path 2d. Paths 2b and 2c involve the path ind60_dem60 (also omitted) so are not further analyzed. These six focus variables are highlighted on the right side of Figure 4.18. From this figure we can see that this subset of variables behaves differently depending on the metric and level of decrement used to define the FPEs. The following sections refer to this figure in analyzing the effects of model misspecification.

For AIC = 2 conditions the dem60_dem60 and dem65_y11 variables remain relatively stable. In contrast, there are larger changes in FPE ranges for the other four focus variables (i.e., dem65_dem60, dem60_y7, dem65_ind60, and y3_y11), which show increasingly large changes that are dependent on the measurement condition. The y3_y11 parameter stands out because it demonstrates the largest relative increase in stability across

parameters. This increase is only the case for the low measurement quality condition.

For RMSEA conditions the smaller decrement condition RMSEA = .001 is more similar to the AIC = 2 condition. The RMSEA = .001 condition variables indicate both increased and decreased stability. However, in the RMSEA = .01 condition variables show exclusively larger FPE ranges when misspecified.

4.2.4.9 FPE Range and FPE Global Range. The previous results have utilized the FPE range metric. This section adds an additional measure designated as FPE global ranges. The FPE global range metric is a more exploratory approach then the more conservative FPE range metric. To first recognize how this set of FP ranges differ from the past set it is useful to contrast the two measures. Traditional FP ranges are calculated as the average maximum FP range within a dataset. The most distal FPEs (i.e., the minimum and maximum estimates) are used to calculate the maximum range within a replication. The mean value of this range across replications is in turn the FPE range metric. The global range treats the 120 replications as a *single* dataset. The distance from minimum and maximum FPEs from this larger dataset constitute the FPE global range. This secondary method is exploratory and has potential disadvantages (notably the comparison of FPEs between replications).

FPE global ranges are shown for the perfect and misspecification 1 conditions for RMSEA (Table 4.4), and AIC (Table 4.5) and for perfect and misspecification 2 conditions (Tables 4.6 and 4.7). RMSEA tables also present both model decrement conditions. In these four tables, the range variable represents the standard FPE range variable, whereas the min.fpe, and max.fpe, represent the minimum and maximum values across all replications within a condition. The seven parameters with the largest FPE global range

values for the misspecified conditions were selected. That is, if a variable had a large (top seven) global range in the misspecified condition it was selected, irregardless of the global range for the same variable in the perfect condition. For RMSEA, selection of top parameters was combined across model decrements. Because the RMSEA decrement = .01 is a larger decrement in fit, the selection of variables was largely determined by the ranges of variables within the RMSEA = .01 condition.

The results from Tables 4.4 and 4.5 indicated several of the focus variables, discussed above, as containing the largest FPE global ranges. For RMSEA, the direction of results was similar to those shown in Figure 4.17 (however, the focus variable dem65_ind60 is more stable using the global range for the perfect condition). For AIC, 4/7 focus variables displayed the largest global FPE ranges. All three latent variables, displayed increased FPE global ranges. In addition to a large FPE global range, the dem65_ind60 also presented estimates that were of different sign: -0.640 and 0.035.

The results for the misspecified 2 condition are shown in Tables 4.6 and 4.7. For RMSEA, 3/6 focus variables were present in the top seven most extreme global FPE ranges. The direction of results for these focus parameters is the same as is shown in Figure 4.18. For AIC, 3/6 focus variables were present in the top seven most extreme global FPE ranges. These three variables in contrast to the standard FPE range, indicate decreased FPE global ranges for misspecified 2 condition.

4.2.5 Summary of Study Factors

This section subsets the results in order to examine the effect of the manipulated factors on FPE ranges. In a generalized FPE analysis, all parameters are allowed to vary,

	variable	focus.v	Dec.	Model	min.fpe	mle.mean	max.fpe	range	range.g
1	dem60~~dem60	0	.001	Perf.	0.357	1.096	1.041	0.019	0.683
2	dem60~~dem60	0	.001	Mis.1	0.619	1.175	1.024	0.011	0.405
3	dem60~~dem60	0	.01	Perf.	0.586	1.105	1.125	0.058	0.539
4	dem60~~dem60	0	.01	Mis.1	0.117	1.170	1.211	0.129	1.095
5	dem60~ind60	1	.001	Perf.	0.204	0.485	0.767	0.058	0.563
6	dem60~ind60	1	.001	Mis.1	0.228	0.499	0.655	0.048	0.427
7	dem60~ind60	1	.01	Perf.	0.119	0.479	0.764	0.201	0.645
8	dem60~ind60	1	.01	Mis.1	0.076	0.501	0.905	0.323	0.830
9	dem65~~dem65	0	.001	Perf.	0.218	0.737	0.767	0.017	0.548
10	dem65~~dem65	0	.001	Mis.1	0.129	0.726	0.698	0.009	0.569
11	dem65~~dem65	0	.01	Perf.	0.145	0.735	0.791	0.046	0.646
12	dem65~~dem65	0	.01	Mis.1	0.156	0.734	0.865	0.089	0.708
13	dem65~dem60	1	.001	Perf.	0.511	0.819	1.169	0.047	0.657
14	dem65~dem60	1	.001	Mis.1	0.533	0.780	1.180	0.037	0.648
15	dem65~dem60	1	.01	Perf.	0.452	0.817	1.174	0.171	0.722
16	dem65~dem60	1	.01	Mis.1	0.369	0.773	1.192	0.285	0.823
17	dem65~ind60	1	.001	Perf.	-0.562	-0.194	0.097	0.051	0.659
18	dem65~ind60	1	.001	Mis.1	-0.584	-0.173	0.079	0.034	0.663
19	dem65~ind60	1	.01	Perf.	-0.564	-0.187	0.155	0.178	0.719
20	dem65~ind60	1	.01	Mis.1	-0.597	-0.165	0.241	0.291	0.838
21	ind60~~ind60	1	.001	Perf.	0.848	1.193	1.326	0.027	0.478
22	ind60~~ind60	1	.001	Mis.1	0.889	1.203	1.147	0.013	0.258
23	ind60~~ind60	1	.01	Perf.	0.785	1.193	1.333	0.075	0.548
24	ind60~~ind60	1	.01	Mis.1	0.705	1.203	1.549	0.162	0.844
25	y6~~y10	0	.001	Perf.	-0.099	0.161	0.271	0.028	0.369
26	y6~~y10	0	.001	Mis.1	-0.066	0.164	0.269	0.018	0.334
27	y6~~y10	0	.01	Perf.	-0.156	0.163	0.316	0.113	0.472
28	y6~~y10	0	.01	Mis.1	-0.453	0.160	0.354	0.222	0.807

Table 4.4. *FPE Values for Misspecification 1 and Perfect Conditions by Model Decrement (RMSEA)*

rather than 2-3 focus variables in previous research. This presents both opportunities and challenges. Allowing all model parameters to vary gives researchers the ability to examine the stability of all model parameters rather than necessitating a limited search of FPEs. However, the large amount of data produced presents challenges when attempting to summarize the study factors effects at each individual parameter.

Multiple sets of fungible parameter estimates are generated for 35 and 64 parameters

	variable	focus.v	Dec.	Model	min.fpe	mle.mean	max.fpe	range	range.g
1	dem60~~dem60	0	2	Perf.	0.630	1.102	0.904	0.003	0.274
2	dem60~~dem60	0	2	Mis.1	0.622	1.174	0.919	0.003	0.297
3	dem60~ind60	1	2	Perf.	0.282	0.481	0.614	0.034	0.332
4	dem60~ind60	1	2	Mis.1	0.272	0.499	0.647	0.037	0.375
5	dem65~~dem65	0	2	Perf.	0.234	0.733	0.676	0.002	0.441
6	dem65~~dem65	0	2	Mis.1	0.103	0.728	0.814	0.005	0.711
7	dem65~dem60	1	2	Perf.	0.508	0.817	1.003	0.026	0.496
8	dem65~dem60	1	2	Mis.1	0.323	0.776	1.244	0.034	0.921
9	dem65~ind60	1	2	Perf.	-0.388	-0.191	0.112	0.029	0.500
10	dem65~ind60	1	2	Mis.1	-0.640	-0.168	0.436	0.035	1.076
11	ind60=~y3	1	2	Perf.	0.505	0.581	0.874	0.078	0.369
12	ind60=~y3	1	2	Mis.1	0.510	0.567	0.779	0.077	0.269
13	y5~~y7	0	2	Perf.	-0.003	0.237	0.247	0.012	0.250
14	y5~~y7	0	2	Mis.1	-0.020	0.238	0.250	0.012	0.270

Table 4.5. *FPE Values for Misspecification 1 and Perfect Conditions by Model Decrement (AIC)*

for the low, and high complexity conditions, respectively. In order to discern these effects, this study conducted an analysis using three methods of output summary. In the overall results section, the mean FPE range of these 35 or 64 parameters is reported as a single value. The results by parameter type subgroup summarizes the parameters according to whether they are classified as covariance, indicator or regression variables for a total of three values per condition (i.e., 162 conditions*3). The final section allows for exploring the results by individual parameter (i.e., minimum of 162*35 parameter ranges per condition).

The first two levels of summary (overall, and by parameter) identified four different interactions that might be investigated (i.e., model condition*fit decrement, model complexity*fit decrement, fit decrement*parameter type (RMSEA), and sample size*model complexity (AIC). The following sections summarizes predictions and results by study factor.

	variable	focus.v	Dec.	Model	min.fpe	mle.mean	max.fpe	range	range.g
1	dem60~~dem60	1	.001	Perf.	0.357	1.096	1.041	0.019	0.683
2	dem60~~dem60	1	.001	Mis.2	0.919	1.000	1.137	0.007	0.218
3	dem60~~dem60	1	.01	Perf.	0.586	1.105	1.125	0.058	0.539
4	dem60~~dem60	1	.01	Mis.2	0.618	1.000	1.583	0.389	0.964
5	dem65~~dem65	0	.001	Perf.	0.218	0.737	0.767	0.017	0.548
6	dem65~~dem65	0	.001	Mis.2	0.265	0.471	0.653	0.004	0.388
7	dem65~~dem65	0	.01	Perf.	0.145	0.735	0.791	0.046	0.646
8	dem65~~dem65	0	.01	Mis.2	0.105	0.469	1.003	0.243	0.898
9	dem65~dem60	1	.001	Perf.	0.511	0.819	1.169	0.047	0.657
10	dem65~dem60	1	.001	Mis.2	0.533	0.717	0.890	0.031	0.357
11	dem65~dem60	1	.01	Perf.	0.452	0.817	1.174	0.171	0.722
12	dem65~dem60	1	.01	Mis.2	0.303	0.718	1.139	0.412	0.836
13	dem65~ind60	1	.001	Perf.	-0.562	-0.194	0.097	0.051	0.659
14	dem65~ind60	1	.001	Mis.2	-0.257	-0.066	0.133	0.045	0.391
15	dem65~ind60	1	.01	Perf.	-0.564	-0.187	0.155	0.178	0.719
16	dem65~ind60	1	.01	Mis.2	-0.476	-0.065	0.339	0.413	0.815
17	ind60~~ind60	0	.001	Perf.	0.848	1.193	1.326	0.027	0.478
18	ind60~~ind60	0	.001	Mis.2	0.911	1.000	1.150	0.009	0.239
19	ind60~~ind60	0	.01	Perf.	0.785	1.193	1.333	0.075	0.548
20	ind60~~ind60	0	.01	Mis.2	0.608	1.000	3.491	0.451	2.883
21	y2~~y2	0	.001	Perf.	0.321	0.467	0.612	0.015	0.292
22	y2~~y2	0	.001	Mis.2	0.274	0.449	0.600	0.008	0.326
23	y2~~y2	0	.01	Perf.	0.283	0.467	0.656	0.056	0.373
24	y2~~y2	0	.01	Mis.2	0.019	0.449	0.774	0.227	0.755
25	y5~~y7	0	.001	Perf.	-0.069	0.238	0.286	0.029	0.355
26	y5~~y7	0	.001	Mis.2	-0.035	0.139	0.287	0.020	0.322
27	y5~~y7	0	.01	Perf.	-0.154	0.239	0.345	0.125	0.499
28	y5~~y7	0	.01	Mis.2	-0.536	0.139	0.430	0.338	0.966

Table 4.6. *FPE Values for Misspecification 2 and Perfect Conditions by Model Decrement (RMSEA)*

4.2.5.1 Sample Size. It was predicted that larger sample sizes would result in smaller FPE ranges. In addition, because AIC is heavily dependent on sample size, it was also predicted that AIC would be more sensitive to sample size increases than RMSEA, which is meant to measure model fit (i.e., epistemic uncertainty) rather than uncertainty due to sample size. These predictions were confirmed by the results in Appendix B2, and the higher proportion of variability in FPE ranges accounted for by Eta-squared for AIC

	variable	focus.v	Dec.	Model	min.fpe	mle.mean	max.fpe	range	range.g
1	dem65~~dem65	0	2	Perf.	0.234	0.733	0.676	0.002	0.441
2	dem65~~dem65	0	2	Mis.2	0.253	0.469	0.618	0.002	0.365
3	dem65~dem60	1	2	Perf.	0.508	0.817	1.003	0.026	0.496
4	dem65~dem60	1	2	Mis.2	0.581	0.718	0.863	0.027	0.282
5	dem65~ind60	1	2	Perf.	-0.388	-0.191	0.112	0.029	0.500
6	dem65~ind60	1	2	Mis.2	-0.217	-0.065	0.113	0.032	0.330
7	y1~~y1	0	2	Perf.	0.406	0.514	0.593	0.005	0.187
8	y1~~y1	0	2	Mis.2	0.223	0.503	0.604	0.007	0.381
9	y3~~y11	1	2	Perf.	-0.023	0.198	0.791	0.013	0.814
10	y3~~y11	1	2	Mis.2	-0.020	0.096	0.254	0.009	0.274
11	y5~~y7	0	2	Perf.	-0.003	0.237	0.247	0.012	0.250
12	y5~~y7	0	2	Mis.2	-0.035	0.138	0.258	0.013	0.294
13	y6~~y10	0	2	Perf.	-0.042	0.158	0.215	0.012	0.257
14	y6~~y10	0	2	Mis.2	-0.061	0.098	0.218	0.013	0.279

Table 4.7. *FPE values for Misspecification 2 and Perfect Conditions by Model Decrement (AIC)*

conditions.

4.2.5.2 Model Complexity. It was anticipated that the high complexity condition would result in smaller FPE ranges than the low complexity condition for RMSEA indexed conditions. The results were mixed. When collapsed across RMSEA decrement conditions, the high complexity condition was most stable (see 4.10). However, when subset by model fit decrement, the larger RMSEA decrement (worse model fit) indicated that the low complexity condition was more stable. No specific prediction was made for AIC indexed conditions which were uniformally more stable for the high complexity conditions.

4.2.5.3 Model Condition. Both models were chosen because they represented relatively moderate levels of model misfit. There was no specific prediction as to what the pattern of FPE would result from model misspecification.

Model Misspecification 1: While 3/7 focus variables and all three latent variables made up the top seven parameters with the largest global FPE ranges the direction of the effect

was dependent on the model decrement. For the smaller decrement size (RMSEA = .001) the misspecified conditions contained smaller ranges, while the larger decrement size (RMSEA = .01) contained larger FPE global ranges for the misspecified conditions. AIC indexed conditions indicates 3/7 focus variables have the largest FPE global ranges. Here, all three of the latent variables show an increase in instability as a result of omitting the y3_y11 variance.

Model Misspecification 2: For RMSEA indexed conditions, the results were dependent on the level of model fit decrement. In general, for the smaller decrement size (RMSEA = .001) the misspecified conditions contained smaller range, although this pattern is less uniform than the misspecification 1 condition (compare Figures 4.15 and 4.16). The larger decrement size (RMSEA = .01) contained larger FPE ranges for all parameters in the misspecified 2 condition. AIC indexed conditions indicate that most variables were more stable for the perfect condition. This result was, however, dependent on whether the standard FPE range or the FPE global range was used. Whereas the standard FPE range indicates more stability for the perfect model the global FPE range metric indicates the misspecified-2 condition as more stable for the focus variables.

4.2.5.4 Measurement Quality. Low measurement quality can sometimes obscure poor model fit in the structural portion of the model. Thus, it was predicted that structural paths would be most stable when measurement quality was low. This pattern was not found uniformly across conditions. For the first misspecification condition, all three FPE ranges were reduced for RMSEA = .001 (see Figure 4.17). The measurement quality was not consistent in predicting the size of FPE range. However, the pattern reverses for the RMSEA = .01 condition with all three latent variable FPE ranges increasing. All three

latent variables had the largest range for the low measurement quality.

The pattern is less clear for the misspecified 2 condition. In this condition, one of the three latent variables is omitted. Once again, RMSEA = .001 indicates the two latent variables' FPE ranges shrinking relative to the perfect condition, and the RMSEA = .01 condition indicating increased FPE ranges relative to the perfect condition.

4.2.5.5 FPE Index and Fit Decrement. It was anticipated that FPEs defined by AIC, and RMSEA, would be differentially affected by sample size. No other explicit predictions were made regarding the differences between RMSEA, and AIC defined indices. Overall, the FPE ranges indicated that the results of the AIC indexed FPEs were closer aligned to the RMSEA = .001 condition (i.e., the smaller decrement condition).

Overall, the set of results indicates a complex relationship between factors, and the stability of parameters as measured by the range of FPEs. The most important factors in predicting FPE ranges are the index of model fit and the size of model fit decrement. The sample size is important for both indexes, but is more important for predicting ranges for AIC indexed parameters. The effect of model misspecification (i.e., model condition) also differentially affected the stability of latent variables (see global fpe range tables). However, the results differed based on type of misspecification, index used, and model decrement. The effect of model complexity was a larger factor AIC indexed conditions. For AIC conditions, high complexity was almost uniformly more stable when compared to the low complexity condition. For RMSEA indexed conditions, the direction of stability was once again dependent on the size of the model fit decrement (see Figure 4.11). The effect of measurement quality was the least important single factor for RMSEA overall (in determining the stability across all model parameters), and second least important single

factor for AIC. For AIC, the FP ranges were largest for latent variables under the misspecified conditions when measurement quality was poor. The effect for RMSEA indexed conditions were, once again, dependent on the size of the model fit decrement.

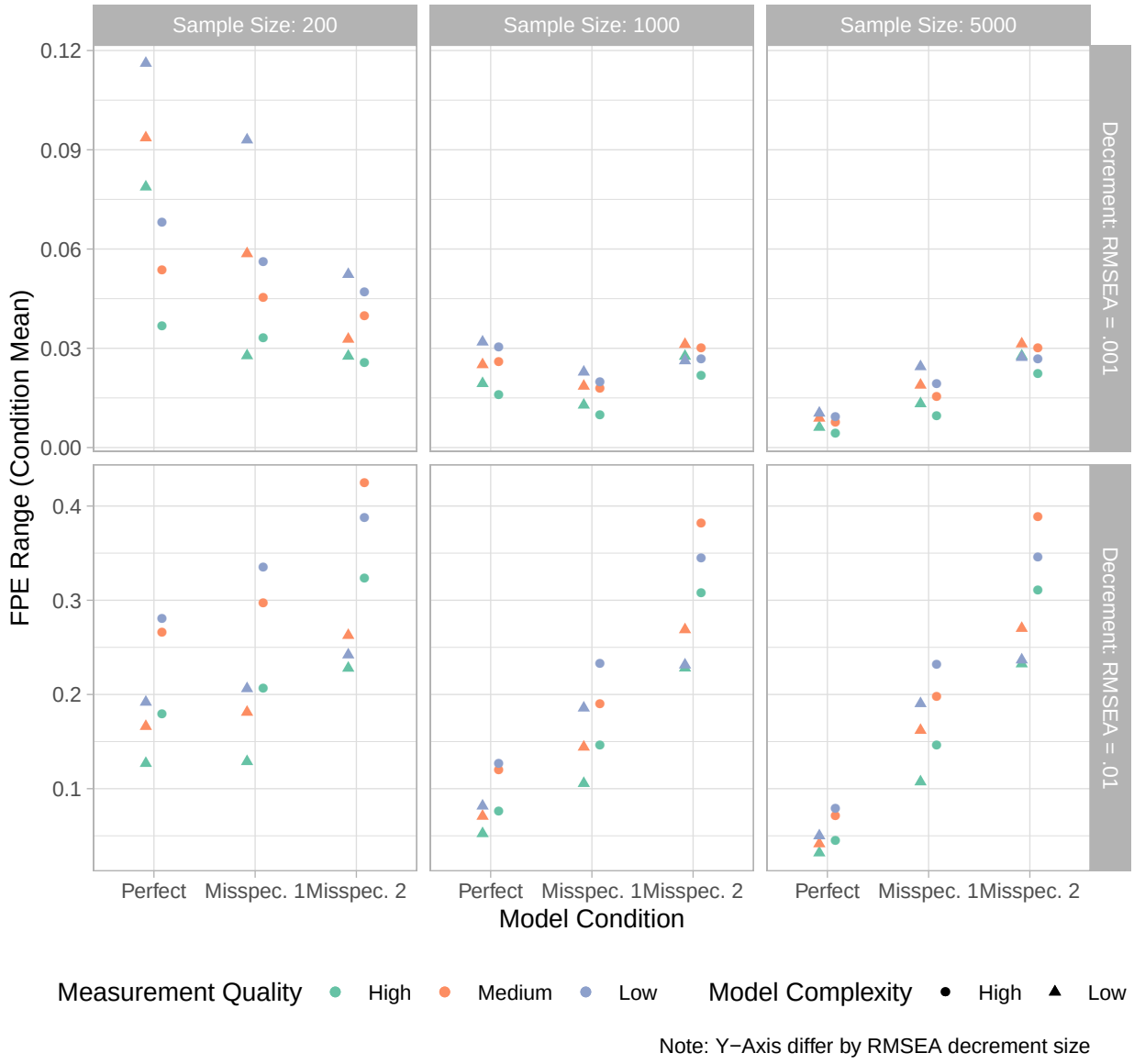

Figure 4.5. FPE range by study factor (RMSEA). Each point represents the mean FPE range averaged across study condition for the conditions that used RMSEA as the index of model fit.

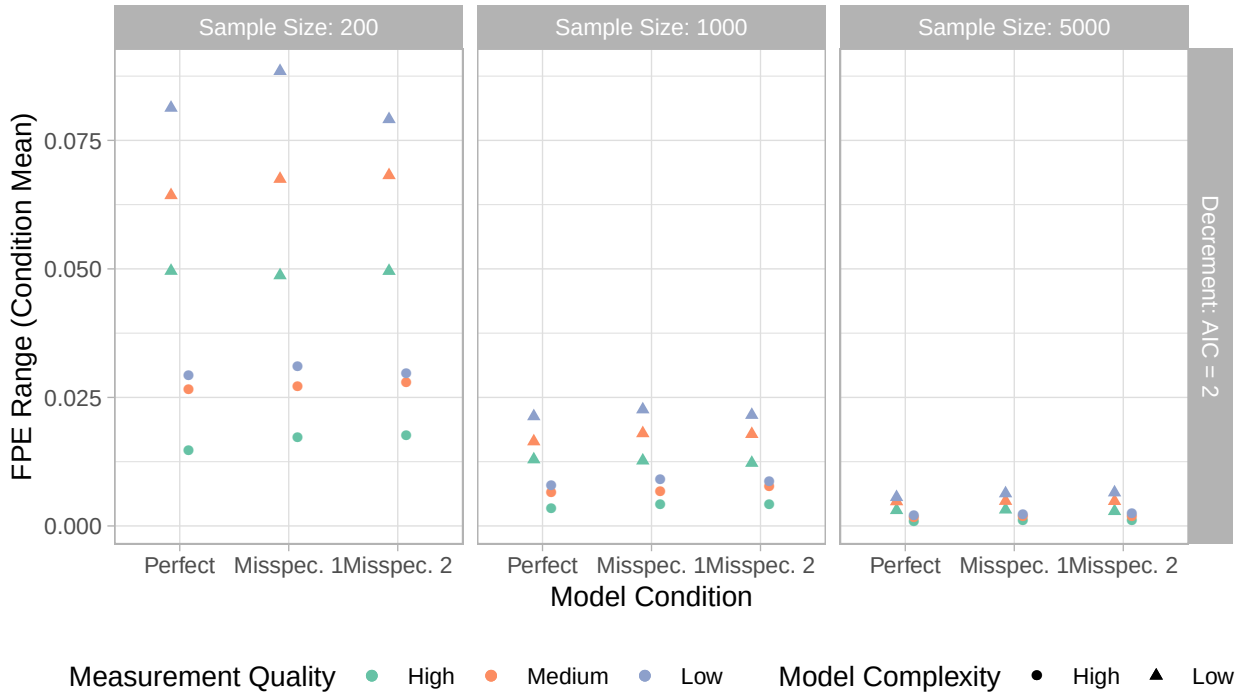

Figure 4.6. FPE range by study factor (AIC). Each point represents the mean FPE range averaged across study condition for the conditions that used AIC as the index of model fit.

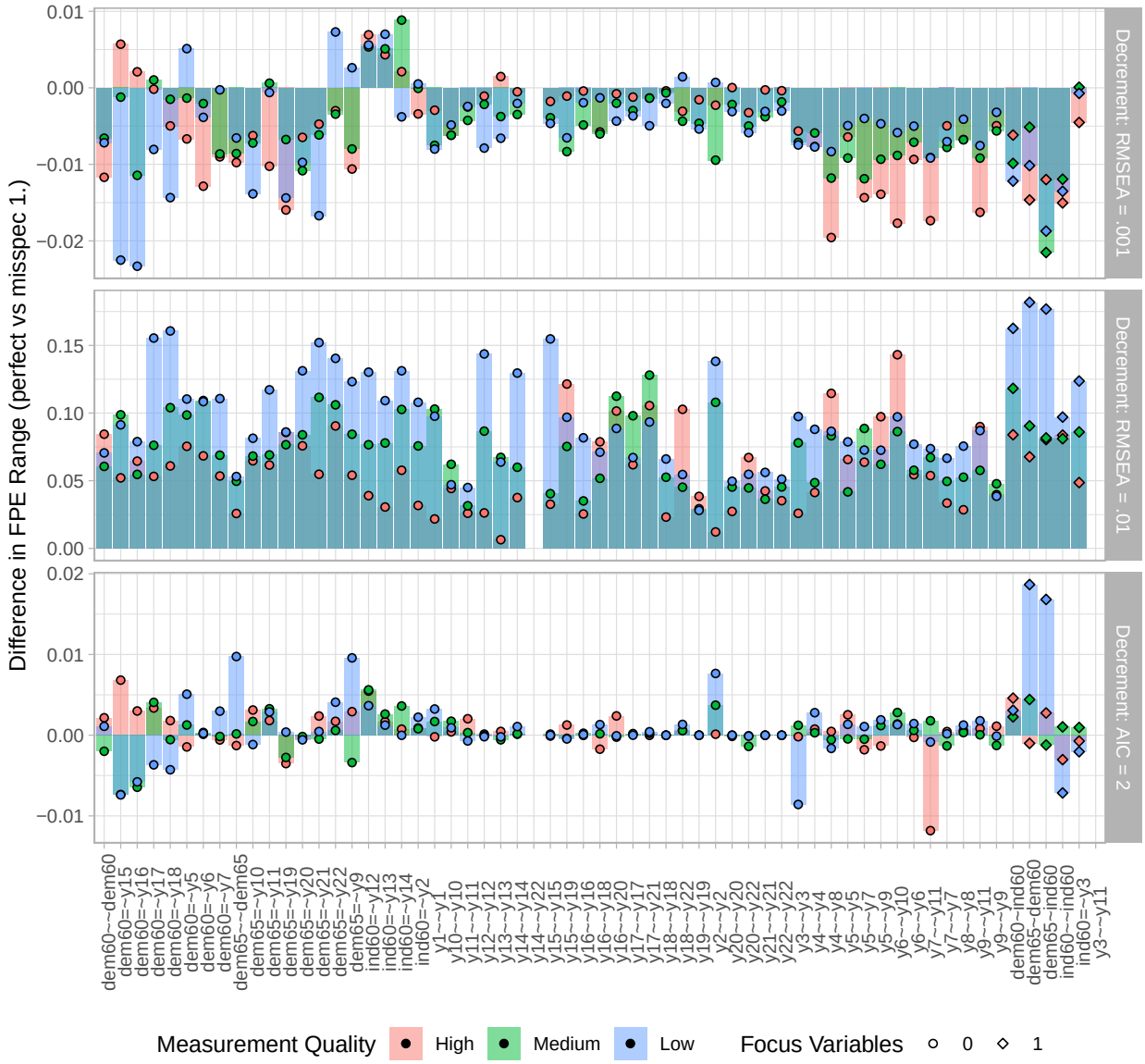

Figure 4.17. Measurement quality and model misspecification 1. The y-axis shows the difference in FPE ranges between the perfect model condition and the misspecified 1 condition. y = 0 represents the value of the perfect model condition and the three deviations represent the difference in model fit under the high, medium and low measurement quality conditions. Negative values represent instances in which the misspecified condition has smaller FPE ranges than the perfect condition whereas positive values are those in which the misspecified condition has larger ranges. The x-axis represents model variables. Conditions with no change values represent those that are not included in both model conditions.

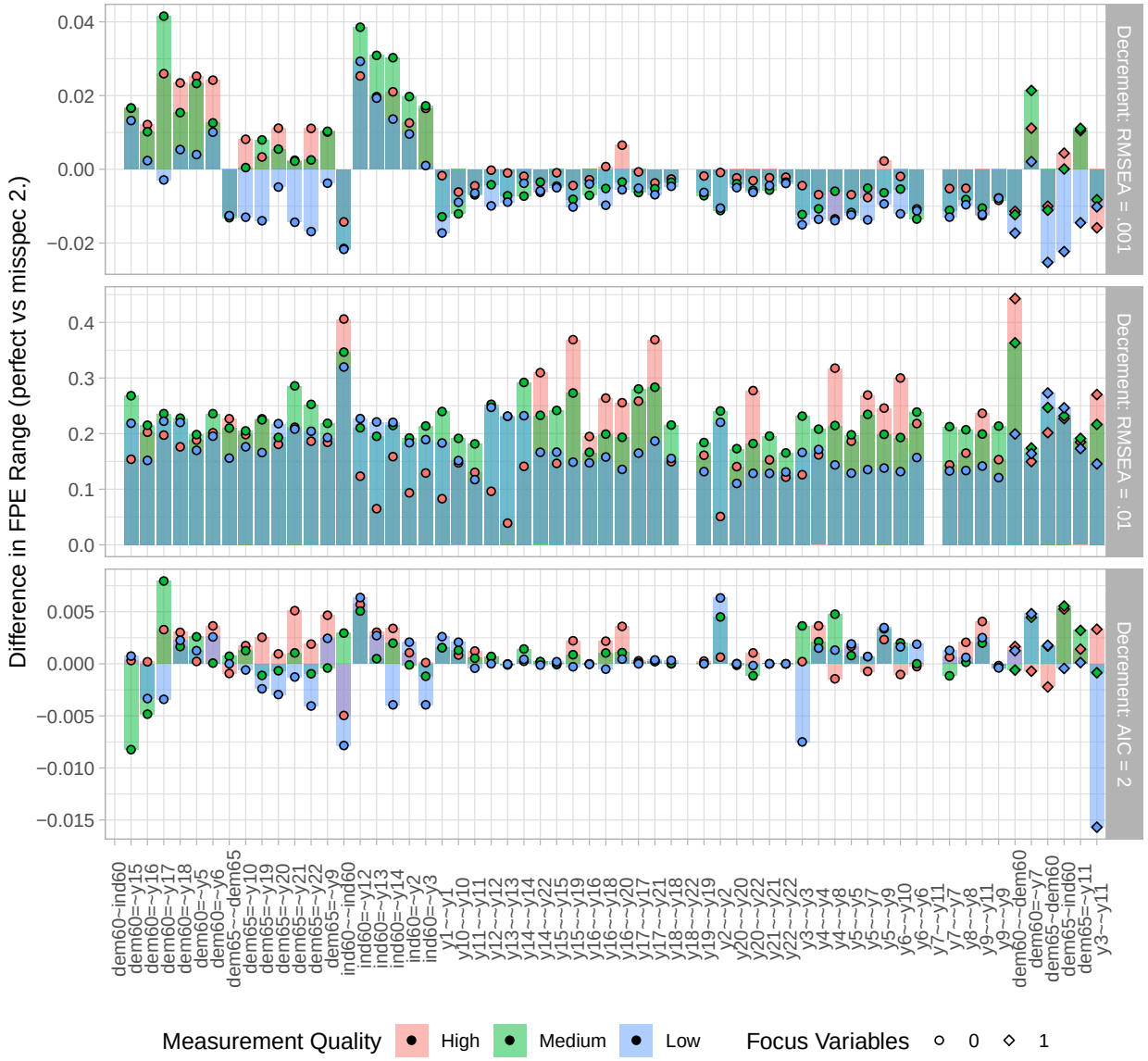

Figure 4.18. Measurement quality and model misspecification 2. The y-axis shows the difference in FPE ranges between the perfect model condition and the misspecified 2 condition. y = 0 represents the value of the perfect model condition and the three deviations represent the difference in model fit under the high, medium and low measurement quality conditions. Negative values represent instances in which the misspecified condition has smaller FPE ranges than the perfect condition. Positive values are those in which the misspecified condition has larger ranges than the perfect condition. The x-axis represents model variables. Conditions with no change values represent those that are not included in both model conditions.

Chapter 5: Discussion

Previous research has demonstrated that alternative estimates can be computed that represent data approximately as well as their maximum likelihood counterparts. The close fit between parameters make these alternative estimates fungible. However, FPEs are in some cases different enough that they would support alternative substantive interpretations of the results when compared to their respective MLEs. This study utilized a method for investigating uncertainty for all parameters simultaneously that differs from previous approaches that used a focal parameter approach. Because the nature of FPEs are not well understood, only exploring FPEs for a few parameters is not justified when the uncertainty of these parameters as compared to other (non-focal) parameters is not known. Conditions and factors encountered in real world data analytic scenarios that may affect the shape of the likelihood function have been largely uninvestigated. To that end, this study investigated the effect of sample size, model complexity, model misspecification, fit decrement, and measurement quality across two different indexes of model-data fit, RMSEA and AIC. The results indicated the presence of a strong interaction between the size of model fit decrement and model misspecification. Latent variable stability was dependent on the size of the fit decrement and the FPE index utilized.

The remainder of this chapter is broken up into several sections. The first section

focuses on the results of the Monte Carlo simulation—drawing conclusions about both main effects and interactions of the manipulated conditions on the stated outcomes. This section then explores their implications to research and the interpretation of FPE results. In a subsequent section, the results will be discussed as test of the FPE framework as a statistical technique. Next the current study results are then compared to previous research. In a final section several study limitations are discussed, and some suggestions for future research given.

5.1 Summary of Results and Implications

The causes of uncertainty relayed by large and small FPE ranges must be understood before a practical analysis of those FPE can be undertaken. That is, inferences regarding fungible parameter estimates can only be used to the extent that researchers understand the causes of *seemingly* informative results (i.e., in this case the range of FPEs).

The potential for interactions between factors on FPE ranges made it necessary that this study considered multiple factors simultaneously. As FPEs are calculated by locating points on the surface of the likelihood function, any factor which can affect the shape of this surface should be considered. The results of this study indicate that there are indeed interactions between factors which necessitated the inclusion of multiple manipulated factors. Important also was the inclusion of two misspecified models and one non-misspecified model conditions. Many, if not most, models are to some degree misspecified. Thus, understanding how misspecifications are manifested in terms of individual parameter uncertainty is necessary if one is to interpret those uncertainties in a meaningful way. This is relevant for researchers who might take a global generalized FPE analysis approach (i.e., FPEs are calculated for all parameters) or a focal parameter

approach.

For the misspecified 1 conditions indexed by AIC, omitting the covariance parameter y3_y11 resulted in large fungible parameter ranges for the focus variables which included the three latent regression parameters. Importantly, these parameters were among the most unstable when compared to other parameters within the model. Misspecification of the model did not uniformly increase the FP ranges across all parameters. This knowledge was previously unknown to those researchers who took a focal parameter approach by generating FPEs for only a small proportion of the total model variables (i.e., FPEs generated for 2-3 parameters). This is interesting because the focus variables were chosen because they were likely to be affected by the omission of the y3_y11 variable (see section 4.2.4.8). This would suggest that the increase in epistemic uncertainty from the model misspecification is realized as a non-random and measurable increase in FPE ranges in several of the latent variables. The insight that these parameters increased in stability when compared to others allows for potentially stronger inferences about how much support there is for the ML estimates over other alternative estimates. In this example, the focus variables also include several latent variables that would likely be of substantive interest to researchers.

To understand these results, one might consider the circumstance in which they are encountered and subsequently could be interpreted in practice. In an applied setting, researchers do not know whether their model is misspecified or if they should be justified in interpreting the results. For example, imagine a scenario in which a researcher is interpreting the results from the misspecified 1 model for the political democracy data set. The researcher conducts an FPE analysis with the goal of determining whether there is

sufficient support for interpreting the ML estimates over alternative explanations (i.e., FPEs). The researcher then interprets the results of FPEs generated using the AIC = 2 model fit decrement. The global FPE ranges for several latent parameters occupy several of the largest FPE ranges. The dem60_ind60 MLE = 0.499 (min.fpe = .272, max.fpe = .646), and dem65_dem60 MLE = 0.776, (min.fpe = .323, max.fpe = 1.244), and dem65_ind60 MLE = -0.168 (min.fpe = -.640, max.fpe = .035). The ranges vary notably from the MLEs for these three latent variables. These fungible parameter estimate ranges do not indicate where a potential misspecification is necessarily located. They do, however, point to alternative estimates that are potentially troublesome for researchers who are interested in interpreting these specific parameter estimates as they represent other valid and interchangeable descriptions of the data set. Whether these points are seen as troublesome depends on whether the fungible estimates are dissimilar enough compared to the MLE to support alternative substantive interpretations of the data. For the political democracy data set, the dem65_ind60 represents the effect of industrialization in 1960 on political democracy in 1965. Whether this effect is -0.640, 0.035, or -0.168 would likely make a difference in the interpretation of the overall results. If the researcher were to add in the omitted y3_y11 covariance and once again calculate FPEs using the same AIC = 2 criteria, they will realize results indicating smaller FPE ranges across the same set of variables (i.e., dem60_ind60, dem65_dem60, dem65_ind60). That is, the second correctly specified model points to smaller FPE ranges.

The results by sample size indicate that FPE ranges for AIC indexed FPEs are strongly influenced by sample size. Furthermore, the AIC = 2 results represent a combination of both aleatory and epistemic uncertainty that is manifested in a seemingly large range for

the dem65_dem60 parameter. As discussed previously, epistemic uncertainty is uncertainty due to the lack of understanding of the true model and the underlying causal processes, whereas, aleatory uncertainty is due to probabilistic uncertainty (often measured by standard errors). Measures of model fit such as RMSEA are thought to be primarily measures of epistemic uncertainty. The results of this study indicate that AIC indexed parameters are strongly influenced by sample size, with larger sample sizes decreasing the ranges of AIC indexed FPEs. Thus the sample size, and model misspecification conditions demonstrate some of the important differences between the two indices of model fit. A researcher interested in interpreting the dem65_dem60 parameter is likely interested in uncertainty that originates from both from modeling uncertainty and insufficient sample size. Irrespective of the origin of uncertainty, the "peakedness" of the likelihood function (i.e., level of support for the MLE) relayed through the range of the FPEs allows researchers to gauge to what extent alternative estimates are equally supported for their particular data set. Though important, sample size is not the only difference between AIC and RMSEA indexed FPEs. For example, there is no difference in sample size between the results shown by RMSEA and AIC in Tables 4.4 and 4.5, yet the results are not equivalent (see also AIC in Figure 4.13, compared to RMSEA in Figures 4.15 and 4.16).

The conclusions might, however, be different if the researcher had instead used RMSEA as the fit measure by which to judge parameter fungibility. For example, the results at the RMSEA = .001 condition indicate the following set of ranges. The dem60_ind60 MLE = .499 (min.fpe = .228, max.fpe = .655), and dem65_dem60 MLE = .780, (min.fpe = .533, max.fpe = 1.180), and dem65_ind60 MLE = -.173 (min.fpe = -0.584, max.fpe = 0.034). These results also included relatively large ranges that might change the substantive

interpretation of these latent variables when compared to the MLEs. If the researcher were to include the previously omitted y3_y11 covariance and re-estimate the FPEs using the same RMSEA = .001 criteria the results would indicate increased uncertainty. That is, using the FPE RMSEA = .001 criteria, the FPE ranges indicate more uncertainty for the perfect model when compared to the misspecified model. In this case, the researcher who does not know the true model might use the information from the FPE analysis to conclude that the misspecified model is preferrable to the perfect condition. At the larger decrement condition, the results indicate the opposite scenario – the misspecified condition is once again indicated as being less stable than the perfect condition. This interaction between the size of the RMSEA model fit decrement, model misspecification and the resulting FPE range is one of the most notable and surprising findings of this study.

Results also indicate that the effect of model complexity (i.e., the number of indicator variables per latent factor) is also dependent on the decrement of RMSEA chosen. Increasing the number of indicators results in larger FP ranges for the RMSEA = .01 condition, but smaller FP ranges for the RMSEA = .001 condition. AIC also revealed smaller FP ranges when more indicators were included.

Lastly, the results by measurement quality additionally demonstrated an unexpected pattern of results. The overall results section indicated that the low and high measurement conditions produced the largest and smallest FP ranges, respectively (see Figure 4.4). The overall results section was made as a summary of the FP ranges across all parameters. When the effect of the factors is, however, not uniform across parameters, this type of analysis is likely not appropriate. It is this individual parameter set of results (Figure 4.17) which are more relevant. However, upon considering model misspecification, the level of fit

decrement, and individual parameters, the results became less easy to parse. For AIC = 2, and RMSEA = .01 (the larger decrement condition) the low measurement quality conditions presents the largest increases in FP ranges when the model is misspecified. This effect of measurement quality does not extend to the RMSEA = .01 condition. These results help to shed light on how FPEs might be used in practice, and what future research might be done to further understand FPEs.

5.1.1 Testing the FPE Framework

Statistics provide the tools to test (i.e., falsify) differing hypothesis and are necessary for descriptive, causal and predictive uses. For example, in psychometrics (a descriptive use of statistics) the goal is often to measure the level of ability of an individual in a given domain. These methods discriminate between the hypothesis represented by differing levels of theta (e.g., low and high ability). Low difficulty items (passed by most) or those too difficult (passed by few) are of limited use because they provide little information as to which hypothesis is most plausible—the level of theta. Instead, items that have a plausible chance of being correct or incorrect given the scenario are ideal. For statistics aimed at assessing a causal hypothesis the goal is similar—discrimination between plausible rival hypothesis. Structural equation modeling is a method of understanding and discriminating between causal hypotheses. Typically, model fit is a popular tool to accomplish this task. If the model (hypothesis) is incorrect (as judged by the index of model fit), then a model might be justifiably rejected. Models that, however, have acceptable fit are often in turn interpreted as reasonable representations of the data and true causal processes. The more difficult the test (as measured by the number of ways it might have failed) the more

impressive the test. SEM results are thus given legitimacy by the testable nature of model fit indexes and the seemingly large number of ways in which a model might have failed—but did not. Model fit, however, is not always effective at discriminating between different models. There are often many equivalent, or nearly equivalent models that might have also been deemed as possessing acceptable fit (MacCallum et al., 1993). These models—while close or identical in model fit—might lead to very different substantive interpretations of the phenomenon under study. Discriminating between disparate outcomes is therefore needed. The promise of an FPE analysis is that it represents a potential additional method of falsifying hypothesis at the level of parameter estimates. Varrying levels of fungible ranges suggest that there is more support for interpreting some parameters than others. Research by MacCallum et al. (2012), and Pek and Wu (2018), and the fictitious researcher examining the political democracy dataset (see example in section above Summary of Results and Implications) have interpreted large FPEs of latent parameters as worrisome. While these results would give research pause and concern there was little research that justified what might influence the range of FPEs in complex latent variable modeling scenarios.

The degree in which one should rely on a statistical technique is connected to the knowledge of what conditions and factors might affect the interpretation of that techniques output. RMSEA is a popular tool to assess model fit in part because of the work by Browne and Cudeck (1992), and, Hu and Bentler (1999) established conditions that affected model fit and suggested guidelines by which it might be interpreted (e.g., RMSEA = .06 fit threshold). If this simulation and work by others had not been conducted, then there would be little basis for interpreting RMSEA. Importantly, researchers have since

pointed to several circumstances in which the RMSEA = .06 threshold for model fit is not appropriate when different levels or factors were considered (Feinian Chen, Curran, Bollen, Kirby, & Paxton, 2008). The current study is similarly necessary to establish a basis for what factors influence the range of FPEs so that they might be interpreted by future researchers. This research indicates many circumstances that complicate the interpretation of FPE ranges. For instance, it was expected that the non-misspecified model would generate smaller FPE ranges than the misspecified model. This was however not consistently the case and was largely dependent on the index used to define the FPEs and the specified level of decrement. This is especially troublesome for researchers who might have used FPE ranges as evidence for preferring one model over another candidate model. This is because within the RMSEA = .001 condition the misspecified 1 model has smaller FPE ranges for several key latent parameters than does the perfect model. The preference towards supporting the perfect model was however found by the AIC = 2, and RMSEA = .01 conditions. FPE ranges may be useful for model selection, or for interpreting a specific models. These results however point that additional research is needed regarding the effect of model misspecification. Because AIC was only studied at one level of model decrement it is unknown if AIC indexed FPEs would continue to point towards the perfect model as being more stable or if the pattern would also reverse at a larger fit decrement (e.g., AIC = 4). The results of this research can also be compared to previous work.

5.1.2 Extensions and Contributions to Research

This study explored factors that influence FPE ranges. To do so several factors were manipulated in a simulation study that were thought to represent scenarios found in

applied settings. The results in this study indicate the complicated relation between these factors and FPE ranges. Given this, what should the interpretation of unstable parameters be (i.e., large fungible ranges)? Past researchers have cautioned against the interpretation of parameters that are unstable based on work from a set of related methodologies: for regression (Wainer, 1976, 1978; Waller, 2008; Waller & Jones, 2009), for canonical correlation analysis (DeSarbo, Hausman, Lin, & Thompson, 1982), and more recently for a structural equation modeling context (T. Lee & MacCallum, 2015; Pek, Chalmers, & Monette, 2016). In each of these instances, the authors warn against the interpretation of optimal estimates when relatively large changes to these estimates result in very similar levels of fit (i.e., flat response surface). However, this research and suggestions are based almost exclusively on real-data examples on which FPEs were calculated. Thus a limited empirical basis for understanding of the causes of parameter stability were available in which to base these recommendations. There were a few examples that did use manipulate factors that may affect the range of FPEs, however. How do these results compare to the current study? Most importantly, Pek et al. (2016) explored sample size, model fit, and the magnitude of the correlation between variables (for a total of 36 conditions). These variables were explored in an effort to distinguish between standard errors (aleatory uncertainty) and fungible parameters (epistemic uncertainty) but nevertheless should be compared to the current results. The present study displayed a similar effect of decreasing FPE ranges with increasing sample size, overall. Pek's study utilized n = 200, 1000, and a population covariance matrix in which to generate FPEs (the current study utilized n = 200, 1000, and 5000). The manner in which model fit conditions is utilized differs between the two studies. Pek's study utilized a method by in which a discrepancy function is

created that meets a particular pre-specified level of model misfit (using only a single model specification). This approach while useful in many contexts, including how misfit affects FPE ranges more generally, it doesn't allow for an consideration of how model misspecifications may lead to differing patterns of FPE ranges or what those patterns might indicate. The current study uses three different model specification conditions to gauge the effect on FPE ranges throughout the model. This is perhaps less generalizable to other models but it is, nonetheless, more realistic in that model misfit in practice is generated largely by model misspecification. In other words, when an investigator interprets a model fit of RMSEA = .04 it does not indicate that their is general misspecification but instead indicates that there are precise underlying misspecifications that do exist–variable omission, path omission, or the functional form is incorrect, for example. Moreover, this fact is more relevant because FPE analysis is not an omnibus methodology (compare to RMSEA, AIC) but instead its level of analysis is focused on the individual parameter. Thus it is a relevant aim to gain insight into how specific parameter instabilities relate to specific misspecifications. This specification approach led to the counter-intuitive finding that several latent variable parameters for the perfect model were at times more unstable than when the model was misspecified. The measured variable correlation manipulation can also be compared to conditions within the present study. The effect of increasing unique variances of the measured variables can be compared to the low reliability conditions within the present study. Pek et al. (2016) indicated that increasing unique variance was associated with large fungible ranges for the focal parameters. Similarly the current study indicated that FPE ranges for latent factors are the smallest for the high measurement quality condition, and the largest for low measurement condition. However, the current

study revealed that the direction of effect of measurement quality also depended on the misspecificaiton condition, and the level of model fit decrement (i.e., the size of the perturbation), and the index in which FPEs were defined (RMSEA vs. AIC). There were, also several other contributions to the existing literature. A suggestion by previous researchers to explore the effects of model complexity had not been previously considered (T. Lee et al., 2017; Pek & Wu, 2018). T. Lee et al. (2017) had suggested that multiple levels of model fit decrement should be examined during an analysis of parameter stability. This was the first study to manipulate this parameter in a SEM context. In part, this was suggested because it was unknown what level of decrement would be most appropriate in practice. Two different levels of fit perturbation were used in the current study and a key finding of this research was the previously unknown interaction between the decrement size and other factors including model complexity, and model misspecification.

Lastly, Pek and Wu (2018) also suggested that future research investigate whether a unified measure that takes into account both epistemic and aleatory uncertainty. To this end, AIC was used as an index of fit that quantifies both of these aspects of uncertainty. Because of it's sensitivity to sample size, and use in model selection it was thought to be a useful candidate measure as an index of informative FPE analysis. Researchers are likely interested in both types of uncertainty – as both sources affect the level of support for the MLE over other candidate estimates. The results for AIC indicated that it had the most promising localization of parameter uncertainty in response to model misspecification.

Another comparison of the current study is to that of the work by Agler and De Boeck (2019). This recent work undertook a simulation study that examined the effect of two variables on the fungible ranges of regression parameters.

Their results indicated an $R^2 = .990$ and $R^2 = 1$ of total variability in FPE ranges is obtained by inclusion of one or both of their manipulated factors, respectively. This work utilized a two and three parameter regression model. The authors found that in the two parameter case that 99% of the variance in FPE ranges was due to the absolute value of the correlation of the other predictor variable and the outcome variable (the other variable examined was the variance inflation factor). For a three predictor model, the fungible ranges of the first predictor using the absolute correlation of the other variables and the criterion resulted in an $R^2 = .839$. While a two and three predictor model are relatively simplistic, the results when contrasted with the current study indicate how studies of varying complexity are necessary. The current study adds to the literature most importantly because it shows the importance of interactions in a relatively simple SEM. In addition to the work by Agler and De Boeck (2019) and Pek and Wu (2018) no other known simulation work has explored which factors explained differing FPE ranges.

5.2 Study Limitations and Future Research

Several limitations may have impacted the conclusions of this study. One of the most surprising findings was the interactions between model fit decrement and several other study factors. AIC only included one level of model decrement that was explored (AIC = 2). This precluded any analysis of whether these interactions with model fit decrement would also occur for AIC indexed parameters or whether this was more exclusive to RMSEA indexed FPEs. For these reasons, the amount of significance placed on these results should be restricted until future research determines the extent of the generalizability of the AIC results. Future research might conduct FPE analysis using a different model with several misspecifications. One challenge of this study is the large

amount of data produced and the difficulty that accompanies summarizing it. Future studies might consider ways in which certain factors may be excluded. For instance, researchers may consider focusing on AIC as a measure of model fit decrement and investigate differing levels of fit decrement (e.g., AIC = 1, 2, and 4) on FPE ranges (but exclude other indices).

While the study used a model that was meant to be representative of a SEM that one might use in practice, it is unknown to the extent in which the results from this study may generalize to other datasets and SEMs. There is a large range of model complexity ranging from the simplistic (e.g., two or three observed predictor variables and one observed outcome variable) to the more complex (e.g., growth mixture modeling). As the number of factors and complexity of the model increase, so does the potential sources of uncertainty.

The results of this study indicated the importance of interactions and how they might affect the likelihood function in surprising ways (e.g., model misfit as revealing smaller FPE ranges in some circumstances). These results would otherwise be difficult to predict without having studied them in conjunction. In several cases complex latent models' standard optimization procedures have difficulty converging to a global optimum due to the presence of local minimum[1] (Li, Harring, & Macready, 2014). The underlying complexity of these likelihood functions would likely be realized in the ranges of fungible parameters. This prediction, however, should be confirmed by future research. On the other end of the complexity spectrum, research has suggested it might be possible (in some circumstances) to completely predict the ranges of fungible estimates (Agler & De Boeck,

[1]While latent growth models are not officially supported by the psindex package, models with convergence issues might explore using the option to use the simulated annealing algorithm to locate the global maximum rather than the nlminb function.

2019). The current simulation involved a greater number of factors, and a more complex model than previous simulations by Agler and De Boeck (2019), and Pek and Wu (2018) (as measured by the number of measured variables, and latent factors). However, in the middle of the complexity continuum, the current three-factor linear SEM is still relatively simple and does not consider the myriad of potential sources of uncertainty and their effects on parameter stability.

This study also did not consider alternative methods for assessing whether we should consider estimates as fungible. Instead of using a fixed threshold (AIC = 2, or RMSEA = .01) future researchers could consider reporting the fit decrement level (e.g., RMSEA = .003) in which the parameters become different enough that they would support substantively different interpretations of the data. This might be useful for interpreting FPEs from different classes of models or levels of complexity. Models that aim to understand complex phenomenon should not be avoided as they often aim to answer questions in which decisions must be made regardless of uncertainty (i.e., post-normal science). It is possible that this threshold may not be constant across model types. It is then more important that future research understand how to interpret uncertainty across model types, and complexity that exists irrespective of researchers quantifying it.

5.3 Conclusion

In part, this study explored whether previously given advice – to avoid interpreting large fungible ranges, and to conversely have preference for smaller fungible ranges– was reasonable. To do so, this study simulated conditions that ought to be preferred (i.e., perfect model condition) to those that should not (misspecified models). In many instances, the fungible ranges for the misspecified conditions were smaller then they were

for the non-misspecified model. This study indicated the complexity of the causes of FPE ranges is as varied as the number of factors that influence the underlying likelihood which they describe and demonstrate some of the challenges to creating easy to follow guidelines for interpreting FPEs. This is because the results of FPE analysis are not only dependent on several different factors that commonly vary in real data analysis, but on the presence of interactions between those factors.

It is possible that FPE analysis reveals the limits of making inferences or predictions based on a particular sample. Wainer (1976) argued that "it is a very rare situation that calls for regression weights which are unequal. This is particularly true in the behavioral sciences, in which relative prediction is the most typical kind of problem". While this might be considered a pessimistic viewpoint it may also suggest the need for more instances of real world or testable implications.

Appendix A: Methods Section Pilot Study Results

variable	Sim. Target (H)	estimate	Absolute Difference
dem65_ind60	0.182	-0.164	0.346
dem65_dem60	0.7	0.773	0.073
y7_y11	0.11	0.111	0.001
y3_y11	0.18	0.181	0.001
ind60_y1	0.92	0.92	0
ind60_y2	0.97	0.97	0
ind60_y3	0.87	0.87	0
dem60_y4	0.85	0.85	0
dem60_y5	0.72	0.72	0
dem60_y6	0.72	0.72	0
dem60_y7	0.85	0.85	0
dem65_y8	0.81	0.81	0
dem65_y9	0.75	0.75	0
dem65_y10	0.82	0.82	0
dem65_y11	0.83	0.83	0
dem60_ind60	0.447	0.447	0
y4_y8	0.296	0.296	0
y5_y7	0.27	0.27	0
y5_y9	0.36	0.36	0
y6_y10	0.19	0.19	0
y9_y11	0.34	0.34	0
y1_y1	NA	0.154	NA
y2_y2	NA	0.059	NA
y3_y3	NA	0.243	NA
y4_y4	NA	0.277	NA
y5_y5	NA	0.481	NA
y6_y6	NA	0.482	NA
y7_y7	NA	0.278	NA
y8_y8	NA	0.344	NA
y9_y9	NA	0.437	NA
y10_y10	NA	0.327	NA
y11_y11	NA	0.311	NA
ind60_ind60	NA	1	NA
dem60_dem60	NA	0.8	NA
dem65_dem65	NA	0.489	NA

Table A1. *Verification of Data Generation for Pilot Model (N = 5 Million)*

variable	Sim. Target (H)	estimate (H)	Sim. Target (M)	estimate (M)	Sim. Target (L)	estimate (L)
ind60_y1	0.92	0.918	0.644	0.64	0.46	0.473
ind60_y2	0.97	0.971	0.679	0.685	0.485	0.486
ind60_y3	0.87	0.869	0.609	0.604	0.435	0.439
dem60_y4	0.85	0.851	0.595	0.599	0.425	0.435
dem60_y5	0.72	0.722	0.504	0.504	0.36	0.369
dem60_y6	0.72	0.724	0.504	0.499	0.36	0.357
dem60_y7	0.85	0.853	0.595	0.596	0.425	0.428
dem65_y8	0.81	0.81	0.567	0.573	0.405	0.422
dem65_y9	0.75	0.749	0.525	0.527	0.375	0.375
dem65_y10	0.82	0.819	0.574	0.569	0.41	0.405
dem65_y11	0.83	0.827	0.581	0.584	0.415	0.406
dem60_ind60	0.447	0.441	0.447	0.435	0.447	0.424
dem65_ind60	0.182	-0.166	0.182	-0.162	0.182	-0.206
dem65_dem60	0.7	0.771	0.7	0.776	0.7	0.823
y4_y8	0.296	0.294	0.296	0.29	0.296	0.283
y5_y7	0.27	0.261	0.27	0.264	0.27	0.247
y5_y9	0.36	0.359	0.36	0.36	0.36	0.367
y6_y10	0.19	0.186	0.19	0.193	0.19	0.191
y7_y11	0.11	0.116	0.11	0.114	0.11	0.11
y9_y11	0.34	0.335	0.34	0.338	0.34	0.335
y3_y11	0.18	0.184	0.18	0.175	0.18	0.179
y1_y1	NA	0.156	NA	0.585	NA	0.759
y2_y2	NA	0.057	NA	0.526	NA	0.745
y3_y3	NA	0.245	NA	0.631	NA	0.791
y4_y4	NA	0.275	NA	0.636	NA	0.795
y5_y5	NA	0.477	NA	0.739	NA	0.846
y6_y6	NA	0.474	NA	0.745	NA	0.859
y7_y7	NA	0.271	NA	0.641	NA	0.8
y8_y8	NA	0.343	NA	0.665	NA	0.803
y9_y9	NA	0.437	NA	0.717	NA	0.842
y10_y10	NA	0.328	NA	0.67	NA	0.817
y11_y11	NA	0.315	NA	0.653	NA	0.817
ind60_ind60	NA	1	NA	1	NA	1
dem60_dem60	NA	0.802	NA	0.8	NA	0.79
dem65_dem65	NA	0.487	NA	0.47	NA	0.386

Table A2. *Verification of Data Generation for Pilot Model (high, medium, and low measurement quality)*

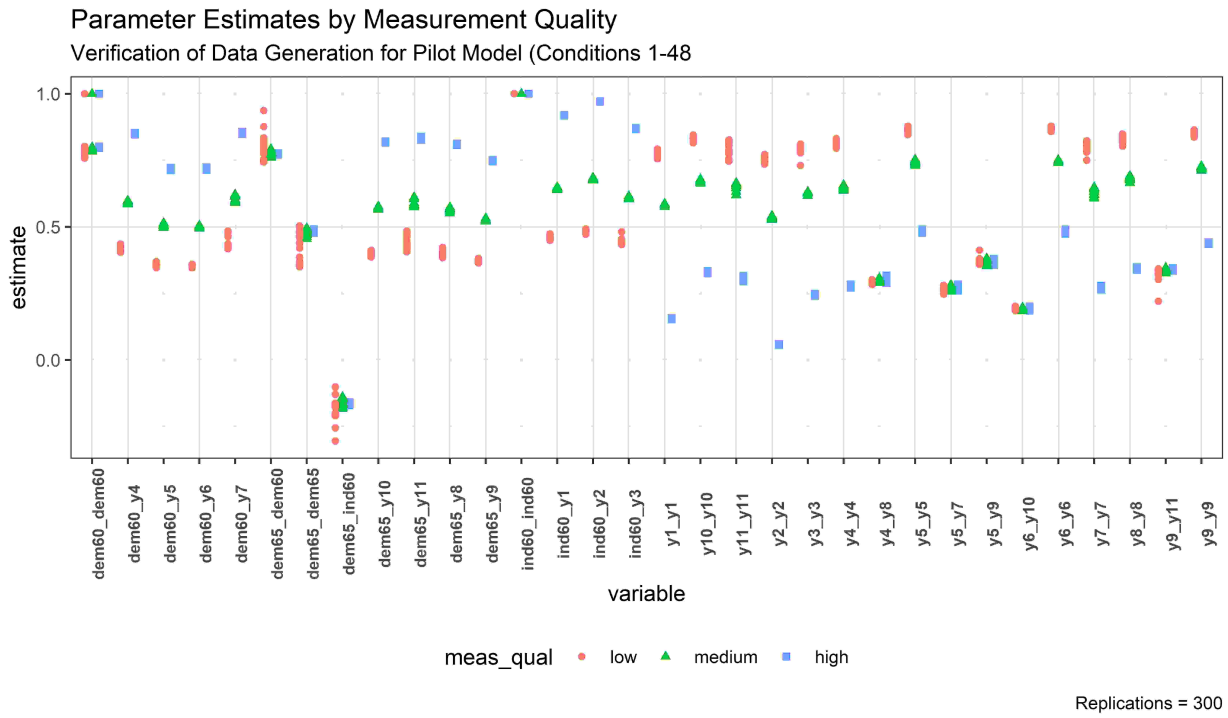

Figure A3. Main Effect of Sample Size by Model Complexity. The upper facets use the FPE Range, while the lower facets use the ln(Percent Change) metric.

Appendix B: Methods Section – Auxiliary Tables and Figures

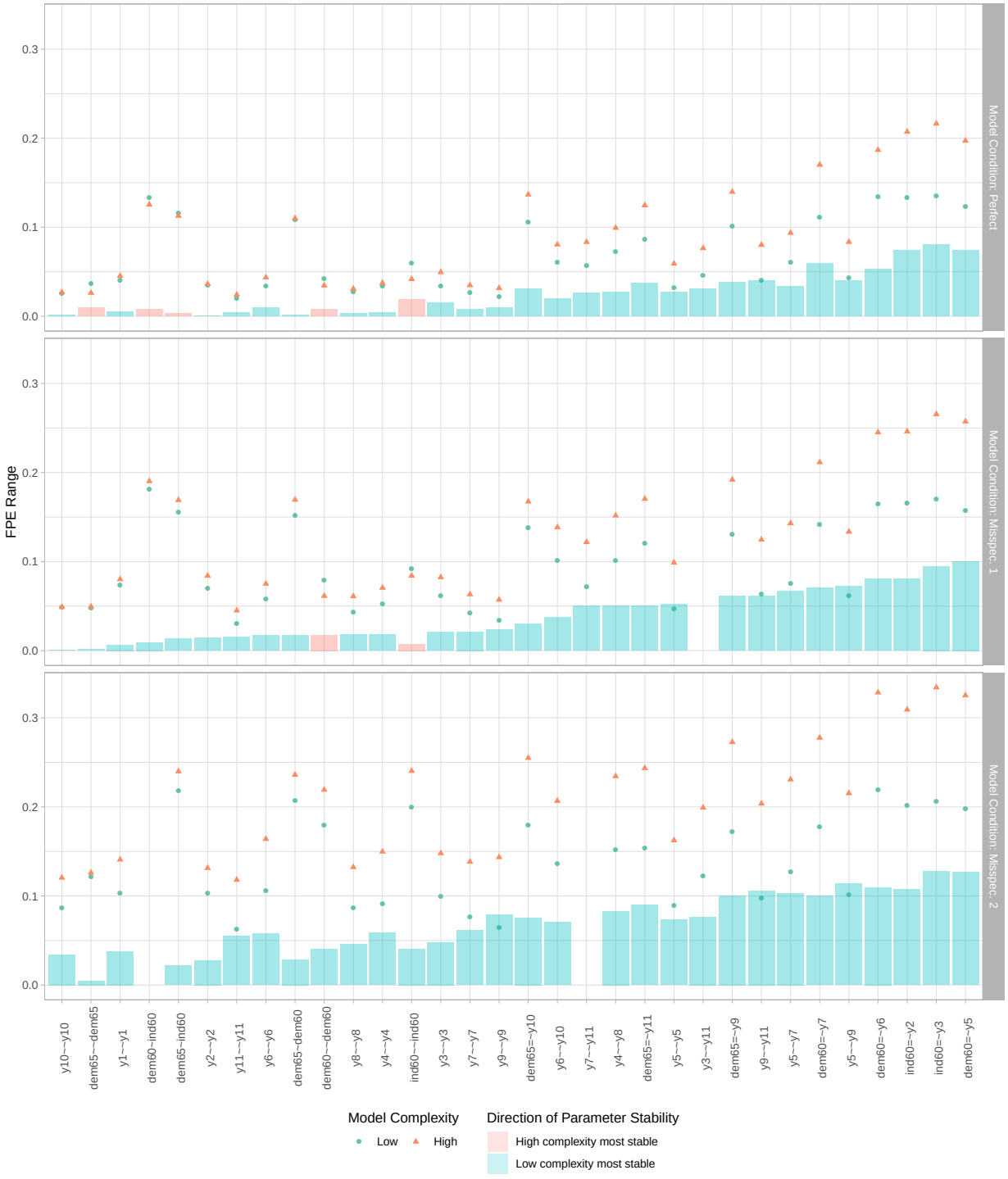

Figure B1. FPE Range for the main effect for model complexity for both AIC and RMSEA. AIC indexed conditions are shown on the top row. The bars are colored blue when the Low complexity condition is relatively more stable than the High complexity condition is more stable than the low complexity condition. Only parameters that are included in both the high and low complexity conditions are included

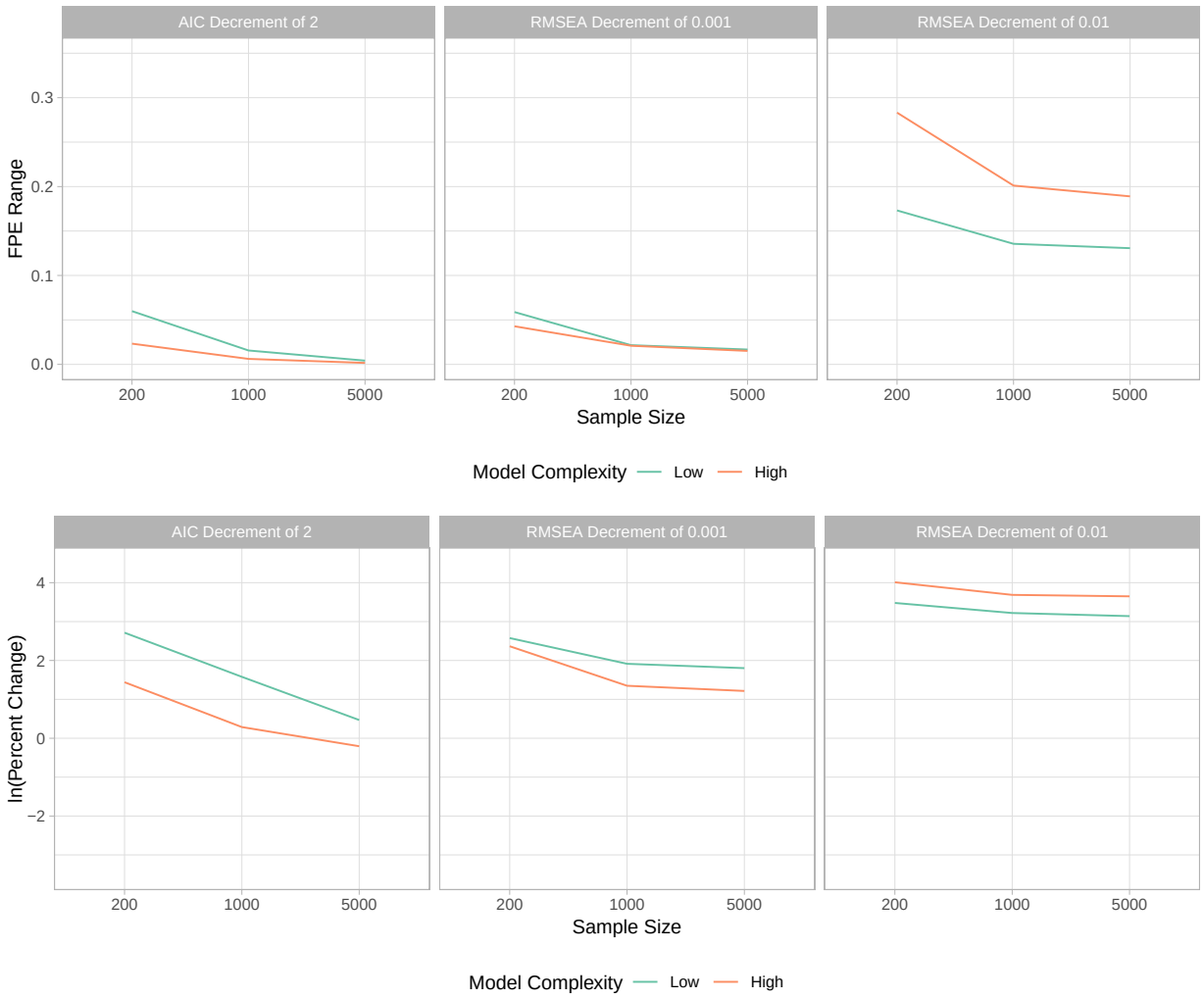

Figure B2. Main Effect of Sample Size (x-axis) shown for FPE range (upper facets) and ln(Percent Change) (lower facets). The two lines represent the different model complexities. The three facets represent the single level of model decrement for AIC and two levels of model decrement for RMSEA.

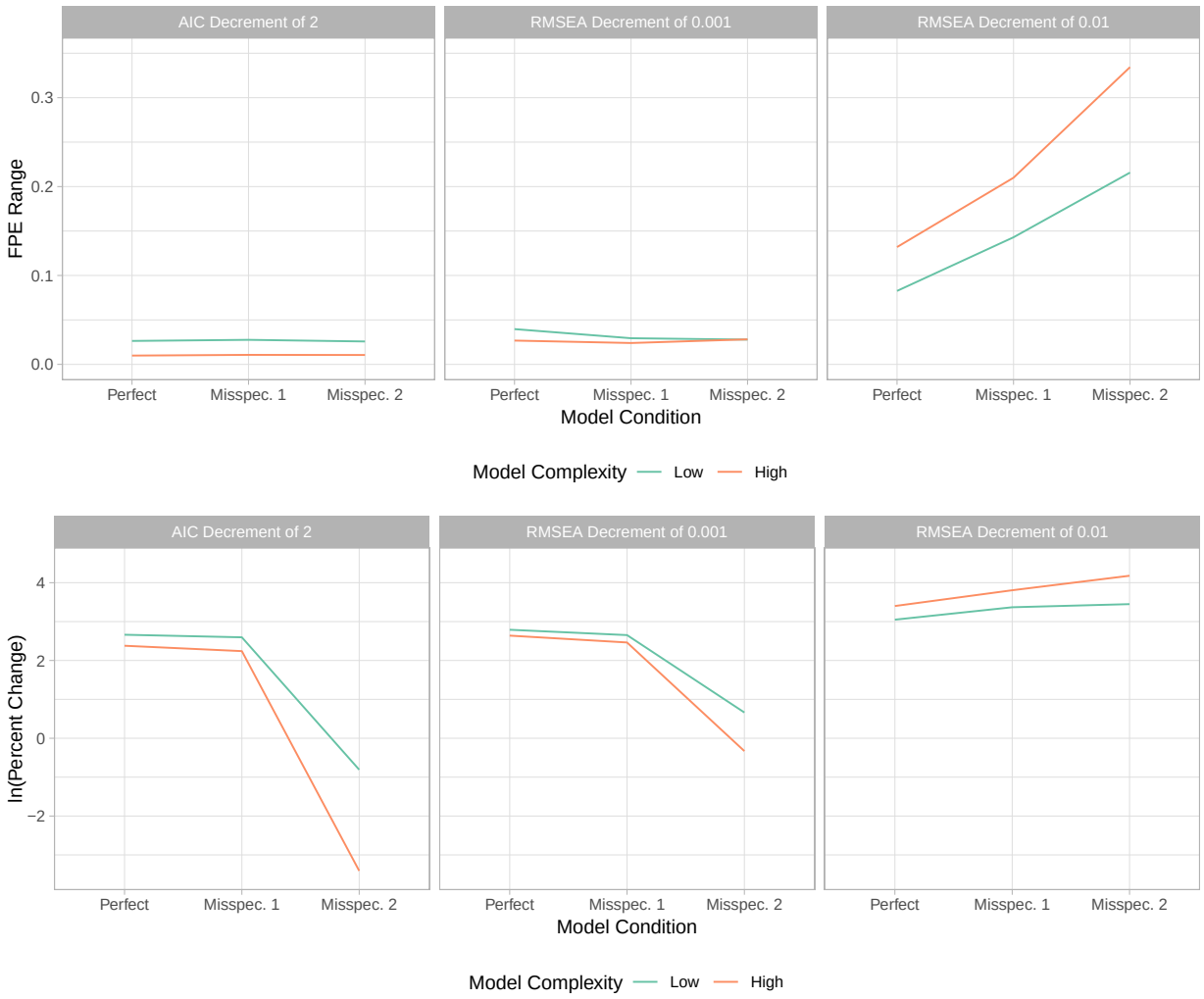

Figure B3. Main Effect of Model Condition (x-axis) shown for FPE range (upper facets) and ln(Percent Change) (lower facets). The two lines represent the different model complexities. The three facets represent the single level of model decrement for AIC and two levels of model decrement for RMSEA.

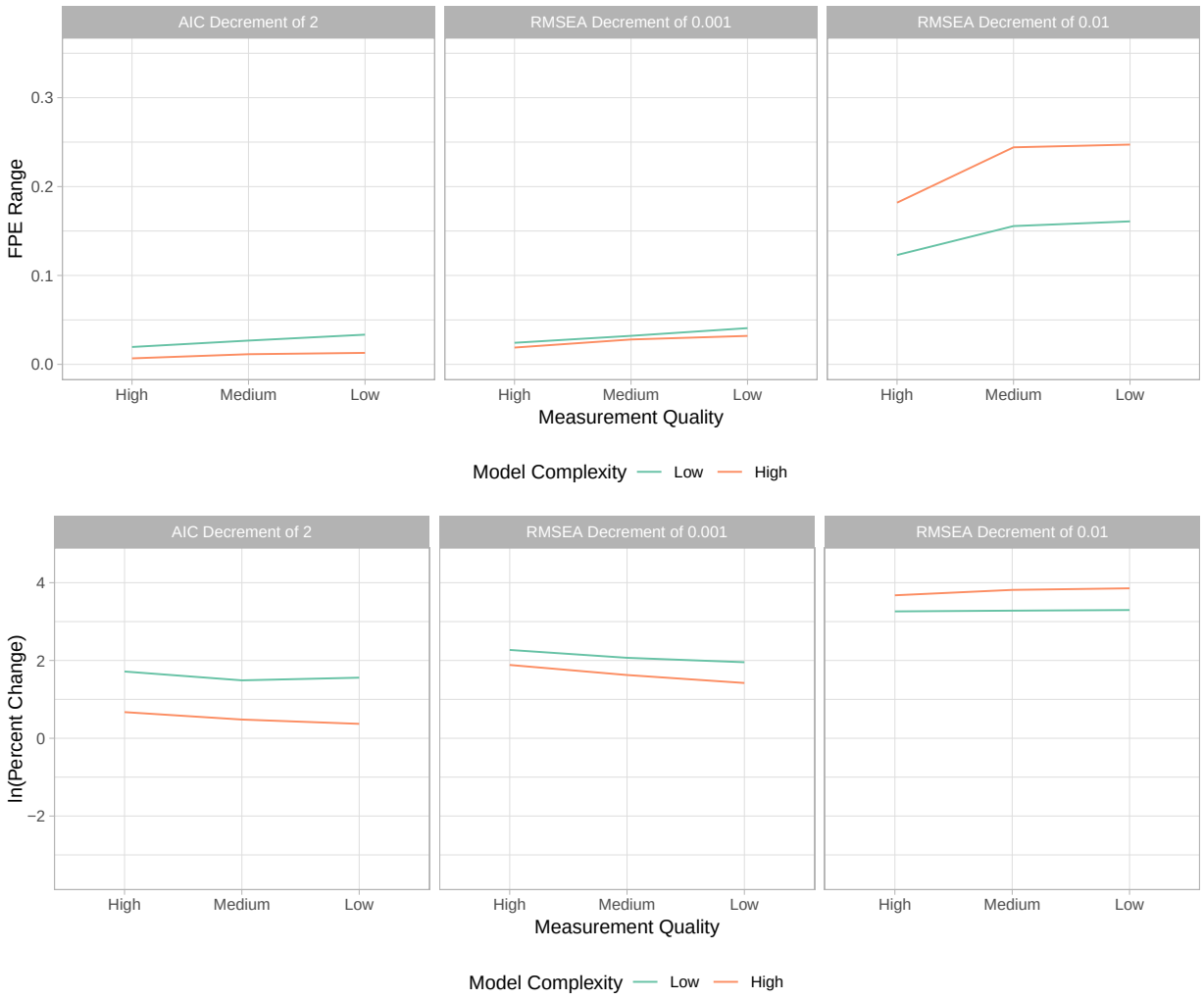

Figure B4. Main Effect of Measurement Quality (x-axis) shown for FPE range (upper facets) and ln(Percent Change) (lower facets). The two lines represent the different model complexities. The three facets represent the single level of model decrement for AIC and two levels of model decrement for RMSEA.

	term	etasq
1	sample_size	0.04
2	‘Model Complexity‘	0.01
3	‘Model Condition‘	0.05
4	measurement_quality	0.01
5	‘Fit Index‘	0.16
6	sample_size:‘Model Complexity‘	0.00
7	sample_size:‘Model Condition‘	0.01
8	‘Model Complexity‘:‘Model Condition‘	0.00
9	sample_size:measurement_quality	0.00
10	‘Model Complexity‘:measurement_quality	0.00
11	‘Model Condition‘:measurement_quality	0.00
12	sample_size:‘Fit Index‘	0.00
13	‘Model Complexity‘:‘Fit Index‘	0.01
14	‘Model Condition‘:‘Fit Index‘	0.02
15	measurement_quality:‘Fit Index‘	0.00
16	sample_size:‘Model Complexity‘:‘Model Condition‘	0.00
17	sample_size:‘Model Complexity‘:measurement_quality	0.00
18	sample_size:‘Model Condition‘:measurement_quality	0.00
19	‘Model Complexity‘:‘Model Condition‘:measurement_quality	0.00
20	sample_size:‘Model Complexity‘:‘Fit Index‘	0.00
21	sample_size:‘Model Condition‘:‘Fit Index‘	0.00
22	‘Model Complexity‘:‘Model Condition‘:‘Fit Index‘	0.00
23	sample_size:measurement_quality:‘Fit Index‘	0.00
24	‘Model Complexity‘:measurement_quality:‘Fit Index‘	0.00
25	‘Model Condition‘:measurement_quality:‘Fit Index‘	0.00

Table B5. $\hat{\eta}^2$ *for all factors representing the manipulated conditions*

	term	$\hat{\eta}^2$
1	‘Fit Decrement‘	0.453
2	‘Model Condition‘:‘Fit Decrement‘	0.097
3	‘Model Condition‘	0.089
4	‘Sample Size‘	0.037
5	‘Model Complexity‘:‘Fit Decrement‘	0.027
6	‘Model Complexity‘	0.018
7	‘Sample Size‘:‘Model Condition‘	0.018
8	‘Measurement Quality‘	0.016
9	‘Measurement Quality‘:‘Fit Decrement‘	0.006
10	‘Model Complexity‘:‘Model Condition‘	0.004
11	‘Sample Size‘:‘Model Complexity‘:‘Fit Decrement‘	0.004
12	‘Model Condition‘:‘Measurement Quality‘	0.004
13	‘Sample Size‘:‘Fit Decrement‘	0.004
14	‘Sample Size‘:‘Model Condition‘:‘Fit Decrement‘	0.003
15	‘Model Condition‘:‘Measurement Quality‘:‘Fit Decrement‘	0.002
16	‘Model Complexity‘:‘Model Condition‘:‘Fit Decrement‘	0.002
17	‘Sample Size‘:‘Measurement Quality‘	0.002
18	‘Sample Size‘:‘Model Complexity‘	0.001
19	‘Model Complexity‘:‘Measurement Quality‘:‘Fit Decrement‘	0.001
20	‘Model Complexity‘:‘Measurement Quality‘	0.001
21	‘Sample Size‘:‘Model Complexity‘:‘Model Condition‘	0.000
22	‘Sample Size‘:‘Model Condition‘:‘Measurement Quality‘	0.000
23	‘Model Complexity‘:‘Model Condition‘:‘Measurement Quality‘	0.000
24	‘Sample Size‘:‘Measurement Quality‘:‘Fit Decrement‘	0.000
25	‘Sample Size‘:‘Model Complexity‘:‘Measurement Quality‘	0.000

Table B6. $\hat{\eta}^2$ *for all variables (RMSEA)*

	term	$\hat{\eta}^2$
1	‘Sample Size‘	0.405
2	‘Model Complexity‘	0.104
3	‘Sample Size‘:‘Model Complexity‘	0.085
4	‘Measurement Quality‘	0.024
5	‘Sample Size‘:‘Measurement Quality‘	0.017
6	‘Model Complexity‘:‘Measurement Quality‘	0.004
7	‘Sample Size‘:‘Model Complexity‘:‘Measurement Quality‘	0.003
8	‘Sample Size‘:‘Model Condition‘:‘Measurement Quality‘	0.000
9	‘Model Condition‘	0.000
10	‘Model Condition‘:‘Measurement Quality‘	0.000
11	‘Sample Size‘:‘Model Complexity‘:‘Model Condition‘:‘Measurement Quality‘	0.000
12	‘Sample Size‘:‘Model Condition‘	0.000
13	‘Model Complexity‘:‘Model Condition‘:‘Measurement Quality‘	0.000
14	‘Model Complexity‘:‘Model Condition‘	0.000
15	‘Sample Size‘:‘Model Complexity‘:‘Model Condition‘	0.000

Table B7. $\hat{\eta}^2$ *for all variables (AIC)*

References

Agler, R. A., & De Boeck, P. (2019, June). Factors associated with sensitive regression weights: A fungible parameter approach. *Behavior Research Methods*. doi: 10.3758/s13428-019-01220-6

Akaike, H. (1974, December). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). New York, NY: Springer New York. doi: 10.1007/978-1-4612-1694-0_15

Azzalini, A. (1996). *Statistical inference: Based on the likelihood*. London: Chapman and Hall.

Beven, K. (2012, February). Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience*, *344*(2), 77–88. doi: 10.1016/j.crte.2012.01.005

Beven, K., & Binley, A. (1992, July). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. doi:

10.1002/hyp.3360060305

Beven, K., & Binley, A. (2014, November). GLUE: 20 years on: GLUE: 20 YEARS ON. *Hydrological Processes*, *28*(24), 5897–5918. doi: 10.1002/hyp.10082

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons, Inc.

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258.

Burnham, K. P., Anderson, D. R., & Burnham, K. P. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (Second ed.). New York: Springer.

Campolongo, F., Cariboni, J., & Saltelli, A. (2007, October). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, *22*(10), 1509–1518. doi: 10.1016/j.envsoft.2006.10.004

Cortez, P. (2014). *Modern Optimization with R*. Cham: Springer International Publishing.

DeSarbo, W. S., Hausman, R. E., Lin, S., & Thompson, W. (1982). Constrained canonical correlation. *Psychometrika*, *47*(4), 489–516.

Edwards, A. W. F. (1972). *Likelihood. An account of the statistical concept of likelihood and Its application to scientific inference*. Cambridge, UK: Cambridge University Press.

Fan, X., Thompson, B., & Wang, L. (1999, January). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 56–83. doi: 10.1080/10705519909540119

Feinian Chen, Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008, May). An

Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models. *Sociological Methods & Research*, *36*(4), 462–494. doi: 10.1177/0049124108314720

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594-604), 309–368.

Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015, June). The Economics of Reproducibility in Preclinical Research. *PLOS Biology*, *13*(6), e1002165. doi: 10.1371/journal.pbio.1002165

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (Third ed., Vol. 2). Boca Raton: CRC press.

Goffin, R. D. (2007, May). Assessing the adequacy of structural equation models: Golden rules and editorial policies. *Personality and Individual Differences*, *42*(5), 831–839. doi: 10.1016/j.paid.2006.09.019

Goldfeld, K. (2017). *Who knew likelihood functions could be so pretty?*

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge, U.K.: Cambridge University Press.

Harring, J. R., McNeish, D. M., & Hancock, G. R. (2017). Using phantom variables in structural equation modeling to assess model sensitivity to external misspecification. *Psychological Methods*, *22*(4), 616–631.

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012, January). Sensitivity of SEM Fit Indexes With Respect to Violations of Uncorrelated Errors. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 36–50. doi:

10.1080/10705511.2012.634710

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.2307/1267351

Hu, L.-t., & Bentler, P. M. (1999, January). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi: 10.1080/10705519909540118

Huang, P.-H. (2017, June). Asymptotics of AIC, BIC, and RMSEA for Model Selection in Structural Equation Modeling. *Psychometrika*, *82*(2), 407–426. doi: 10.1007/s11336-017-9572-y

Jones, J. A., & Waller, N. G. (2016). Fungible weights in logistic regression. *Psychological Methods*, *21*(2), 241–260. doi: 10.1037/met0000060

Kenny, D. A., & McCoach, D. B. (2003, July). Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(3), 333–351.

King, G. (1998). *Unifying political methodology: The likelihood theory of statistical inference*. Ann Arbor: University of Michigan Press.

Koehler, E., Brown, E., & Haneuse, S. J.-P. A. (2009, May). On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. *The American Statistician*, *63*(2), 155–162. doi: 10.1198/tast.2009.0030

Koopman, R. F. (1988). On the sensitivity of a composite to its weights. *Psychometrika*, *53*(4), 547–552. doi: 10.1007/BF02294406

Kuhn, T. S. (1970). *The structure of scientific revolutions* ([2d ed., enl ed.). Chicago:

University of Chicago Press.

Kullback, S., & Leibler, R. A. (1951, March). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. doi: 10.1214/aoms/1177729694

Lagani, V., Triantafillou, S., Ball, G., Tegner, J., & Tsamardinos, I. (2016). Probabilistic computational causal discovery for systems biology. In L. Geris & D. Gomez-Cabrero (Eds.), *Uncertainty in Biology* (pp. 33–73). Cham: Springer.

Lai, K., Green, S. B., & Levy, R. (2017). Graphical displays for understanding SEM model similarity. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(6), 803–818. doi: 10.1080/10705511.2017.1334206

Lee, S.-Y., & Wang, S.-J. (1996). Sensitivity analysis of structural equation models. *Psychometrika*, *61*(1), 93–108.

Lee, T., & MacCallum, R. C. (2015). Parameter influence in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 102–114. doi: 10.1080/10705511.2014.935255

Lee, T., MacCallum, R. C., & Browne, M. W. (2017). Fungible parameter estimates in structural equation modeling. *Psychological Methods*, *23*(1), 58–75. doi: 10.1037/met0000130

Leek, J. T., & Peng, R. D. (2015). What Is the Question? *Science*, *347*(6228), 1314–1315.

Levy, R., & Hancock, G. R. (2011). An extended model comparison framework for covariance and mean structure models, accommodating multiple groups and latent mixtures. *Sociological Methods & Research*, *40*(2), 256–278. doi: 10.1177/0049124111404819

Li, M., Harring, J. R., & Macready, G. B. (2014, May). Investigating the Feasibility of

Using Mplus in the Estimation of Growth Mixture Models. *Journal of Modern Applied Statistical Methods*, *13*(1), 484–513. doi: 10.22237/jmasm/1398918600

MacCallum, R. C., Lee, T., & Browne, M. W. (2012). Fungible parameter estimates in latent curve models. In M. C. Edwards & R. C. MacCallum (Eds.), *Current topics in the theory and application of latent variable models* (pp. 183–197). New York, NY: Routledge.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*(1), 185–199. doi: 10.1037//0033-2909.114.1.185

McNeish, D., An, J., & Hancock, G. R. (2018, January). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment*, *100*(1), 43–52. doi: 10.1080/00223891.2017.1281286

McNeish, D., & Hancock, G. R. (2018, March). The effect of measurement quality on targeted structural model fit indices: A comment on Lance, Beck, Fan, and Carter (2016). *Psychological Methods*, *23*(1), 184–190. doi: 10.1037/met0000157

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140. doi: 10.1037/met0000078

Morris, M. D. (1991). Factorial Sampling Plans for Preliminary Computational Experiments. *Technometrics*, *33*(2), 15.

Moshagen, M. (2012, January). The Model Size Effect in SEM: Inflated Goodness-of-Fit Statistics Are Due to the Size of the Covariance Matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 86–98. doi:

10.1080/10705511.2012.634724

Nash, J., & Sutcliffe, J. (1970, April). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. doi: 10.1016/0022-1694(70)90255-6

Neyman, J., & Pearson, E. S. (1933). IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Phil. Trans. R. Soc. Lond. A*, *231*(694-706), 289–337.

Open Science Collaboration. (2015, August). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-aac4716. doi: 10.1126/science.aac4716

Patil, P., Peng, R. D., & Leek, J. T. (2016, July). What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspectives on Psychological Science*, *11*(4), 539–544. doi: 10.1177/1745691616646366

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001a, April). Monte Carlo Experiments: Design and Implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287–312. doi: 10.1207/S15328007SEM0802_7

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001b, April). Monte Carlo Experiments: Design and Implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287–312. doi: 10.1207/S15328007SEM0802_7

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (Second ed.). Cambridge, UK: Cambridge University Press.

Pek, J. (2012). *Fungible parameter contours and confidence regions in structural equation models* (Doctoral Dissertation). University of North Carolina, Chapel Hill.

Pek, J., Chalmers, R. P., & Monette, G. (2016, October). On the Relationship Between Confidence Sets and Exchangeable Weights in Multiple Linear Regression. *Multivariate Behavioral Research*, 0–0. doi: 10.1080/00273171.2016.1225563

Pek, J., & MacCallum, R. C. (2011, April). Sensitivity Analysis in Structural Equation Models: Cases and Their Influence. *Multivariate Behavioral Research*, *46*(2), 202–228. doi: 10.1080/00273171.2011.561068

Pek, J., & Wu, H. (2015, December). Profile Likelihood-Based Confidence Intervals and Regions for Structural Equation Models. *Psychometrika*, *80*(4), 1123–1145. doi: 10.1007/s11336-015-9461-1

Pek, J., & Wu, H. (2018, December). Parameter uncertainty in structural equation models: Confidence sets and fungible estimates. *Psychological Methods*, *23*(4), 635–653. doi: 10.1037/met0000163

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016, May). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, *79*, 214–232. doi: 10.1016/j.envsoft.2016.02.008

Popper, K. (1959). *Conjectures and refutations: The growth of scientific knowledge.*

Prendez, J. Y., & Harring, J. R. (2019). Measuring Parameter Uncertainty by Identifying Fungible Estimates in SEM. *Journal Structural Equation Modeling: A Multidisciplinary Journal*, 33.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

Ravetz, S. O., & Ravetz, J. R. (1993). Science for the post-normal age. *Futures*, *25*(7),

735–755.

Razavi, S., & Gupta, H. V. (2015, May). What do we mean by sensitivity analysis? The need for comprehensive characterization of "global" sensitivity in Earth and Environmental systems models: A Critical Look at Sensitivity Analysis. *Water Resources Research*, *51*(5), 3070–3092. doi: 10.1002/2014WR016527

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Saltelli, A. (Ed.). (2004). *Sensitivity analysis in practice: A guide to assessing scientific models*. Hoboken, NJ: Wiley.

Saltelli, A. (Ed.). (2008a). *Global sensitivity analysis: The primer*. Chichester, England ; Hoboken, NJ: John Wiley. (OCLC: ocn180852094)

Saltelli, A. (Ed.). (2008b). *Global sensitivity analysis: The primer*. Chichester, England ; Hoboken, NJ: John Wiley. (OCLC: ocn180852094)

Saltelli, A., & Annoni, P. (2010, December). How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, *25*(12), 1508–1517. doi: 10.1016/j.envsoft.2010.04.012

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., ... Tarantola, S. (2008). *Global sensitivity analysis: The primer*. Chichester, England: John Wiley & Sons, Inc.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009, October). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. doi: 10.1080/10705510903203433

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2),

461–464.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Shmueli, G. (2010, August). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. doi: 10.1214/10-STS330

Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, *11*, 1643–1662.

Steiger, J. H. (2000, June). Point Estimation, Hypothesis Testing, and Interval Estimation Using the RMSEA: Some Comments and a Reply to Hayduk and Glaser. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(2), 149–162. doi: 10.1207/S15328007SEM0702_1

Steiger, J. H. (2007, May). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(5), 893–898. doi: 10.1016/j.paid.2006.09.017

Steiger, J. H. (2016, November). Notes on the Steiger–Lind (1980) Handout. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 777–781. doi: 10.1080/10705511.2016.1217487

Steiger, J. H., & Lind, J. C. (1980, May). Statistically based tests for the number of common factors. In *Spring Meeting of the Psychometric Society.* Iowa City, IA.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*(2), 213–217. doi: 10.1037/0033-2909.83.2.213

Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin*, *85*(2), 267–273. doi: 10.1037/0033-2909.85.2.267

Wainwright, H. M., Finsterle, S., Jung, Y., Zhou, Q., & Birkholzer, J. T. (2014, April). Making sense of global sensitivity analyses. *Computers & Geosciences*, *65*, 84–94. doi: 10.1016/j.cageo.2013.06.006

Waller, N. G. (2008). Fungible weights in multiple regression. *Psychometrika*, *73*(4), 691–703. doi: 10.1007/s11336-008-9066-z

Waller, N. G., & Jones, J. A. (2009). Locating the extrema of fungible regression weights. *Psychometrika*, *74*(4), 589–602. doi: 10.1007/s11336-008-9087-7

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37. doi: 10.1037/1082-989X.8.1.16

Wu, H., & Neale, M. C. (2012, November). Adjusted Confidence Intervals for a Bounded Parameter. *Behavior Genetics*, *42*(6), 886–898. doi: 10.1007/s10519-012-9560-z

Xiang, Y., Gubian, S., Suomela, B., & Hoeng, J. (2013). Generalized Simulated Annealing for Global Optimization: The GenSA Package. *The R Journal*, *5*, 16.