Title
Challenges of Evaluating the Causal Effects of Early Child Development Programs

Author
Weber, Ann

Publication Date
2012

Peer reviewed|Thesis/dissertation

Challenges of Evaluating the Causal Effects of Early Childhood Development Programs

By

Ann Weber

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Epidemiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:
Professor Ira B. Tager, Co-Chair
Professor Lia H. Fernald, Co-Chair
Professor Maya Petersen
Professor Mark Wilson

Spring 2012

Abstract

Challenges of Evaluating the Causal Effects of Early Childhood Development Programs

By

Ann Weber

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor Ira B. Tager, Co-Chair
Professor Lia H. Fernald, Co-Chair

Over 200 million children under five years old in low- and middle-income countries (39% of preschool children in developing countries) are estimated as not achieving their potential across multiple domains of development (including sensori-motor, cognitive, language, and social-emotional development). Although there is evidence of benefit to child development for a wide range of interventions, results from assessments of scaled-up programs are less conclusive. Therefore, the assessment of large-scale early child development (ECD) programs in developing countries is a priority. This dissertation focuses on several methodological issues in evaluating large-scale ECD interventions that threaten the validity of finding a program benefit. Specifically, I address two important areas of evaluation: 1) the challenge of obtaining an unbiased measure of language development in a setting for which the test was not developed; and 2) the analytic process of determining whether the ECD intervention had a benefit that actually is the result of the intervention, given an unbiased developmental outcome.

To demonstrate these challenges, I make use of data collected over a 14 year period for a national nutrition program in Madagascar. First implemented in 1999 by the National Office of Nutrition (ONN) in Madagascar, the program has expanded to include 5550 sites with coverage of approximately 1.1 million children. The program takes a comprehensive approach to improving early child nutritional status, targeting children less than 5 years of age and including multiple activities that have been found to be associated with better child outcomes. A wide spectrum of developmental outcomes was assessed in four national surveys in Madagascar, including physical growth (height and weight), and motor, cognitive, language, and behavioral skills. In my dissertation, I focus on only two of these measures: weight-for-age (a measure of short-term nutritional status) and receptive language (understanding of words, gestures or phrases), as assessed by an adaptation of the U.S. version of the Peabody Picture Vocabulary Test, 3rd edition (PPVT-III).

Tests of early child cognition and language that were developed and carefully validated in one country are not guaranteed to maintain their properties when adapted and translated for use in another. The risk of censoring is high, and bias from differential item functioning (DIF) can be introduced when administering the test to different subgroups (e.g., ethnicities) within the same

country. Using longitudinal data from two rounds of testing (when children were 3-6 years and 7-10 years of age) I apply item response theory (IRT) models to assess the performance of the PPVT in Madagascar. My analysis uncovers problem items (e.g., bias from DIF by dialect spoken in 55% of items), censoring in a large proportion of the children, and patterns of responses related to test fatigue. This information can be used to identify items that need to be dropped before estimating the program effect (e.g., items with strong, significant DIF); and items that should be replaced, modified, or re-ordered in future work (to avoid censoring and test fatigue). Although my analysis focuses on a test of vocabulary, many of the issues apply to any multi-item instrument intended to capture a latent construct. Such multi-item measures are commonly used in ECD intervention research and include other tests of language and memory, as well as non-verbal tests of cognition and socio-emotional behavior scales. I present lessons learned from working with the PPVT in Madagascar and make recommendations for how these lessons can be applied in other developing country settings.

Presuming that the developmental outcome is assessed without bias, there remains the analytic challenge of determining whether an ECD intervention has a benefit that actually is the result of the intervention. I make use of a detailed, step-by-step roadmap for estimating the average treatment effect (ATE) of Madagascar's program on children's mean weight-for-age in a community between 1997 and 2004. The evaluation of the Madagascar program is complicated by the fact that the selection into the program was non-random and strongly associated with the pre-treatment (lagged) outcome. The availability of pre-program data allows me to define the outcome as either the post treatment value or the change from pre-treatment to post-treatment. Using these two outcome definitions, I contrast identification results for three common statistical parameters that under different assumptions are equivalent to my target parameter, the ATE. These statistical parameters are a post-treatment estimand commonly used in epidemiology that adjusts for measured confounders, and two difference-in-differences estimands (one of which is popular in econometrics) that can address certain types of unmeasured confounders. For identification, I make the assumptions underlying each of these estimands explicit and demonstrate the consequences of alternate choices using directed acyclic graphs and data simulations. Finally, I describe and compare three methods of estimation for each of the three estimands: traditional parametric regression, inverse probability of treatment weights (IPTW or propensity score weighting), and targeted maximum likelihood estimation (TMLE). Throughout, I avoid imposing parametric model assumptions unless they are firmly supported by knowledge, and deliberately keep the process of identification separate from the process of estimation in order to avoid the common confusion of the two.

My findings show that I am faced with a serious bias trade-off when choosing an estimand for the ATE of the Madagascar nutrition program. A post treatment estimand controls for confounding due to the lagged outcome but not from possible unmeasured confounders. The difference-in-differences estimands do not control for confounding by lagged outcome, but have the potential to adjust for a certain type of unmeasured confounding. However, the difference-in-differences estimands have the potential for introducing bias if the additional assumptions they require (beyond those needed for the post-treatment estimand) are not met. The three estimands result in very different estimates of effect in the Madagascar study, regardless of method of estimation. The estimates for the ATE from the post-treatment estimand are less than one tenth of a standard deviation (SD) improvement in community mean weight-for-age z-score

(as compared to the WHO reference population). The two difference-in-differences estimands are comparable to each other, with estimates of the ATE ranging from 0.24 to 0.28 SD increase in mean weight-for-age z-score (statistically significant with all estimation methods). However, since I am unable to estimate either the magnitude or direction of possible confounding from unmeasured factors, or the magnitude or direction of bias from the failure of the assumptions to hold, I conclude that my best choice is the post treatment estimand. This simple estimand adjusts optimally for known measured confounders and is equal to the ATE under the fewest assumptions.

Given this choice of estimand, the choice of estimator can still make a difference. In fact, the only significant effect for the post-treatment estimand was obtained with TMLE (estimate of the ATE = 0.066 SD, CI: 0.001, 0.146 SD). TMLE has specific advantages over either parametric regression or IPTW, and improves on both by implementing a bias reduction step to estimate the target parameter of interest. In addition, TMLE is considered doubly robust to model misspecifications. Therefore, I conclude that TMLE is a better choice for estimation over the other two methods, and that my best estimate of the ATE is small, but statistically significant. Alternate target parameters and alternate estimation approaches are unlikely to resolve the uncertainty of the choice of estimand for the Madagascar evaluation. However, future analytic work on other nutritional outcomes (e.g., height-for-age) and longer term effects of the program (e.g. from a third wave of data in 2011) may provide an accumulation of evidence of a causal benefit of Madagascar's nutrition program.

There is mixed evidence of the effectiveness of large-scale nutrition programs on early child development outcomes. Although the lack of consistent results is generally attributed to possible problems of implementation and governance of the program, the failure to find a statistically significant effect (or alternatively, the success of finding one) may, in fact, be due to the types of problems described in my dissertation. There may be bias in the outcome (or other) measure, failure of the causal assumptions to hold, or bias from the method of effect estimation. Misleading estimates of a program's benefit (in either direction) have significant policy and funding implications for the program. More importantly, the decisions made based on an evaluation have consequences for the children the programs are trying to help. I present tactics for addressing several methodological challenges to evaluation and urge investigators to update and/or reconsider their analytic approaches to evaluations. Since ECD intervention research is often inter-disciplinary, I recommend learning new methods from other disciplines and to use the best methods that are at our disposal.

Dedication

To

my father, Arnold R. Weber, M.D., M.P.H.

in memory
of his devotion
to his work and to his family,
and for
setting an example of
how to
live a life of adventure and learning

To

my husband, Karl Pope, M.P.H.

for
his love and patience,
his encouragement
and
enthusiasm for my work,
but most of all,
for making the dreams we dream together
come true

To

my family and friends

for believing in me
and for
not giving up when I didn't call, write or visit

Table of Contents

Chapter 1: Background 1

1.1 Introduction 1
1.2 Why a Malnutrition Intervention in Madagascar? 2

1.2.1 A Framework for Early Child Development 3
1.2.2 Madagascar's Comprehensive Program 4
1.2.3 Child Development Outcomes 5

1.3 Measurement – the case of language 6
1.4 Identification of the target parameter – the case of pre/post data 8
1.5 Effect estimation – a comparison of methods 9
1.6 Implications 9

Chapter 2: Measurement – the Case of Language 11

2.1 Introduction 11

2.1.1 Argument for IRT over Traditional 11
2.1.2 Argument for additional DIF testing 12
2.1.3 Argument for Multidimensional model 14
2.1.4 Summary 15

2.2 Methods 15

2.2.1 Sample & Language Data 15
2.2.2 Classical Test Methods 16
2.2.3 Unidimensional Models 17
2.2.4 Multidimensional Model 19
2.2.5 Software 22

2.3. Results 22

2.3.1 Sample Characteristics 22
2.3.2 Separate Unidimensional Models 24
2.3.3 Multidimensional Model 30

2.4 Discussion 33

2.4.1 Test Validity 33
2.4.2 Use of Multidimensional Model 34
2.4.3 Recommendation 34
2.4.4 Conclusions 36

Chapter 3: Identification of the Target Parameter – the Case of Pre/Post Data 38

3.1 Introduction 38
3.2 Methods 39

3.2.1 Setting 39
3.2.2 Notation 39
3.2.3 Causal Inference Road Map 40
3.2.4 Illustration of Results Using Simulated Data 49

3.3 Discussion 51

Chapter 4: Effect Estimation – a Comparison of Methods 54
4.1 Introduction 54
4.2 Methods 56
4.2.1 Data 56
4.2.2 Identifiability in the Madagascar Context 58
4.2.3 Estimation Procedures 60
4.2.4 Statistical Inference 65
4.2.5 Software Packages 65
4.3 Results 65
4.3.1 Exclusion restriction assumptions 65
4.3.2 Checks for Positivity Violations 66
4.3.3 Estimation Results 67
4.4 Discussion 69
4.4.1 Estimation Results 69
4.4.2 Conclusions 72
4.4.3 Future Work 73
Chapter 5: Conclusions 75
5.1 Overview 75
5.2 Measurement – the case of language 75
5.3 Identification of the target parameter – the case of pre/post data 77
5.4 Effect estimation – a comparison of methods 78
5.5 Final Remarks 80
References 81
Appendices 88
A1: Additional Figures and Statistics for IRT Study 88
A2: R Code for Simulations 97
A3: Supplementary Information for Estimation 100

Acknowledgements

This thesis would not have been possible without the mentorship, guidance and support I received from my committee members at Berkeley: Professors Ira B. Tager, Lia H. Fernald, Maya Petersen and Mark Wilson. I am grateful to each of them for investing in my work in meaningful, unique and thoughtful ways.

I am indebted to the many people whose hard work and dedication made the collection of data in Madagascar possible, in particular: Mme Lisy Ratsifandrihamanana and her group of Social Services students; Valerie Ranaivo, Jean Harvel Randriamanjakasoa and staff at the National Institute of Statistics; Jean Rakotosalama and staff at the National Office of Nutrition; Voahirana Rajoela at the World Bank; and the SEECALINE Community Health Workers. I am also indebted to the thousands of children and mothers who donated their time to participate in the surveys.

I am grateful for funding received throughout my graduate studies, including fellowships from the Center for Global Public Health and the Graduate Division. I would especially like to thank Professor Ray Catalano for encouraging me to apply for a Robert Wood Johnson (RWJ) fellowship. The RWJ funding led to a fruitful collaboration with Professor Maya Peterson and resulted in two chapters in my thesis. In addition, I am extremely grateful for the work funded by Professors Paul Gertler and Lia Fernald that preceded my thesis work and precipitated my application to the PhD program.

I would also like to thank other professors at Berkeley who have encouraged me or offered advice during my training: Professors Art Reingold, Amani Nuru-Jeter, Barbara Abrams, and Mahasin Mujahid, as well as the always helpful and friendly staff in the Epidemiology office: Roberta Myers, Eugene Konagaya and Juanita Cook.

I've been inspired and energized by all of the doctoral students in the Epidemiology program and their very important work. It was a pleasure and honor to learn alongside my PhD cohort: Caitlin Gerdts, Joshua Gruber, Raymond Lo, Benjamin Chaffee, Hope Biswas, Aracely Tamayo, Ayse Ercumen, and Ling-I Hsu. I also appreciated sharing ideas with the participants in Professor Fernald's research group: Patricia Kariger, Kallista Bley, Jessica Jones-Smith, Emily Flynn, Melissa Hidrobo, Rachel Gardner, and Christopher Golden.

Finally, I owe my deepest gratitude to Dr. Emanuela Galasso, Senior Economist at the World Bank. Emanuela gave me my first paid job as a data analyst to evaluate the performance of the battery of child development tests administered in Madagascar. She has been generous with the data from Madagascar, allowing me to apply my "epi" methods, contrast them to various econometrics approaches, and then to write about my findings. Although we occasionally agree to disagree, she is forever encouraging of my work.

Chapter 1: Background

1.1 Introduction

Over 200 million children under five years of age in low- and middle-income countries (39% of preschool children in developing countries) are estimated as not achieving their potential across multiple domains of development.[1] These domains are broad and encompass sensori-motor, cognitive, language, and social-emotional development. Key risk factors have been identified that are associated with this failure in achievement, including inadequate cognitive stimulation, iodine deficiency, and other factors frequently associated with poverty (e.g., infections).[2, 3] In a review of the effectiveness of existing interventions that aim to improve early child development (ECD), the authors find evidence of "substantial and positive effects" from a number of interventions studied in experimental settings.[4] However, they report that "results from assessments of scaled-up programmes were more variable."[4] Possible explanations given for this discrepancy include the difficulty of expanding coverage of a program while maintaining quality, or the lack of local capacity-building for implementation. The authors recommend prioritizing the assessment of national ECD programs, but they do not discuss the complexities of planning and performing such an evaluation. The potential limitations and pitfalls of the usual methods involved in evaluating the success (or failure) of large-scale interventions are often overlooked.

Problems exist throughout the evaluation process and are not limited to the significant logistical and technical constraints of administering a survey to thousands of households in hundreds of communities in a national study. This dissertation addresses problems related to two important methodological areas in evaluation: 1) the challenge of obtaining an unbiased measure of cognitive or language development in an ethnically and linguistically diverse low-income country setting; and 2) the analytic process of determining whether the ECD intervention had a benefit that actually is the result of the intervention (i.e., had a causal effect). Given that we are frequently faced with less than ideal data in an evaluation, our ability to make a causal claim of benefit may be compromised. For example, when the intervention is not randomly assigned for ethical reasons (as is often the case in a scaled-up program), investigators are forced to confront the fact that communities that receive the intervention may be systematically different than those that do not (i.e., there is confounding). However, the underlying motivation of an evaluation is to learn if there is a causal benefit of the program. A significant association alone may be insufficient evidence that the program "works." If we want to inform policy based on our research, then we must minimize possible sources of bias and tackle causal inference with all of its complexities. To demonstrate some of these challenges in the evaluation of ECD interventions, I make use of a long-running national nutrition program in Madagascar. Although my work is based on a study conducted in a low-resource setting, the methods and conclusions are applicable to the evaluation of early childhood programs and policies throughout the world.

This background chapter is organized as follows. Section 1.2 explains why a national nutrition program in Madagascar provides an excellent opportunity to explore the complexities of evaluation. Section 1.3 describes the measurement challenge faced by researchers who want to evaluate program effects on a measure of cognitive development, specifically of language ability. The challenges associated with estimating a causal benefit using observational data are split into two sections. Section 1.4 addresses the problems of definition and identification of the target parameter of interest (i.e., the estimate used to infer program benefit or lack thereof), and section

1.5 addresses the problem of estimating the target parameter. Section 6 concludes with the implications of this work.

1.2 Why a Malnutrition Intervention in Madagascar?

Malnutrition has wide-ranging consequences for children's health and development, including their physical growth, gross and fine motor skills, cognitive and language skills, and social-emotional development. [5, 6] Malnutrition is characterized by both under- and over-nutrition, and is typically assessed with the use of binary indicators that compare the weight and height (or length, depending on age) of a child to that of a reference population of well- nourished and healthy children of the same age and gender. The most common indicators for under-nutrition are: underweight, stunting and wasting, which indicate that a child is two standard deviations below the median of the reference population for weight-for-age, height-for-age and weight-for-height, respectively.

Underweight is a near-term marker for inadequate nutrition and is estimated to be responsible for the largest proportion of the death and disease burden associated with malnutrition (approximately 19% of deaths and disability adjusted life years in children under 5 years).[6] Stunting is associated with long-term under-nutrition and has been identified as a key predictor of later cognitive function (discussed in more detail below) with sufficient evidence available to recommend intervention.[1, 7] Poor cognitive development is also associated with severe micro-nutrient deficiencies (e.g., from lack of iodine) independently of the link through stunting.[8, 9] Wasting is an indicator of recent weight loss (e.g., from starvation or disease), and in severe cases is cause for immediate treatment and/or referral to a health clinic.

In Madagascar, approximately 50 percent of children under five are estimated to be stunted.[10] In 1999, the Madagascar National Office of Nutrition (ONN) implemented a program, named SEECALINE, to help address this problem. The project experienced a period of sustained growth throughout the country through 2008, expanding to include 5550 sites with coverage of approximately 1.1 million children (or about a third of children under 5 years of age in Madagascar).[11] The program expansion stopped in 2009 after a presidential coup and subsequent economic crisis, which is unresolved to this day. Despite the recent political and economic instability, the program has remained operational, making SEECALINE one of the few long-running and large-scale nutrition interventions in Africa.[11] The government's continued investment and ownership of the program is evidenced by the integration of the program into long-term strategic priorities of the country and their stated interest in expanding the intervention to include additional activities for participants (e.g., early child stimulation).[12] The ONN hopes that evidence of long term benefits of the SEECALINE program will justify the continuation of the program in the event of renewed international engagement of the donor community.

The Madagascar program provides an opportunity to fill in the gap of knowledge of the effectiveness of nutrition programs running on a national scale in a low-income country. In addition, SEECALINE offers several compelling reasons for study: an emphasis on a sensitive period of child development, a comprehensive approach to improving children's nutritional status, and the availability of nationally representative assessments of a wide range of outcomes. The relevance of each of these aspects of the program to evaluating an ECD intervention will be discussed in the sub-sections below.

1.2.1 A Framework for Early Child Development

The timing of the Madagascar program is strategic: the first five years of life are periods of rapid development when children's developmental trajectories are inextricably linked to influences of their environment (nutritional and otherwise).[13] Advocates of intervening early in a child's life refer to the study of neuroscience and the developing brain for their supporting evidence. Specifically, they refer to the notion of critical and sensitive periods during which time areas of the brain mature and become specialized.[14] From a life-course perspective, these periods can be interpreted broadly to represent a time window during which an exposure can have either adverse or protective effects on development and subsequent outcomes.[15] A critical period represents a limited window, whereas a sensitive period has a wider scope of opportunity for modifying or reversing the effect of exposures beyond a critical time frame. If we intervene during either of these periods, the brain (or body) may have sufficient plasticity (or resilience) to recover from earlier deprivation and achieve normal or near-normal development. For example, the first two years of life are considered a strategic window (i.e., a critical period) for intervening to prevent stunting, which is generally considered irreversible after this point.[16] More recently, however, there has been some evidence of catch-up growth after the age of two years, suggesting a wider window (i.e., a sensitive period).[17]

In contrast to height-for-age, cognitive development is thought to be characterized by a long sensitive period. Cognitive function, or cognition, refers to the developmental domain responsible for processing and applying information, and includes sustained attention, memory, problem solving, decision making, and language proficiency.[18] Although the first few years of life are characterized by children's rapidly increasing ability to interact with others and express themselves (as any parent can attest), higher cognitive processes associated with the prefrontal cortex, such as language and cognitive control, have not achieved adult levels of development until late in adolescence.[14]

Evidence from animal models indicates that specific areas of the brain integral to cognitive development, such as the cortex, hippocampus and striatum, are sensitive to nutritional deficiencies during the first two years of life. [9, 19] In humans, naturally occurring experiments provide some of the best evidence of how sensitive humans are to environmental deprivation and the importance of the timing of an intervention to improve cognitive outcomes. One particularly fascinating example is the English and Romanian Adoptees (ERA) study.[20] Children living in state-run Romanian orphanages experienced varying durations of severe deprivation prior to their adoption into the United Kingdom (UK). Researchers found that Romanian children placed prior to 6 months of age performed as well cognitively at age 4-6 years as children born in the UK. But children placed after 6 months had lower scores on intelligence tests at 4-6 years, as compared to children placed before 6 months.[20] In addition, the investigators found evidence of a long sensitive period for cognitive development. Adopted children with the lowest IQ scores at age 6 years (98% of whom were over 6 months on arrival in the UK) significantly increased their score by age 11 (although they did not catch up completely with higher scoring children).

By targeting nutrition during the first five years of life, Madagascar's SEECALINE program works within a framework for child development that addresses a critical and sensitive period model. In addition, the program may have a latent effect on children. Specifically, a life-course latency model suggests that exposure to an intervention during early periods of development can

have effects that influence outcomes later in life (even decades later).[15, 21] This model and other life course models (i.e., cumulative and pathway models) provide a framework for the link between early stunting and later cognitive outcomes. Evidence for such a link is primarily found in observational studies and natural experiments. The strongest evidence comes from a meta-analysis of 5 longitudinal studies in low to mid-income countries. The authors of the analysis found that stunting in children less than 2 years of age was predictive of shorter adult height, lower attained schooling, reduced adult income, and decreased off-spring birth-weight.[16] Strong evidence also comes from natural experiments, in particular the English and Romanian Adoptees (ERA) study discussed above.[20] In addition, cognitive abilities established in childhood have been shown to persist into adulthood: skills in pre-school are associated with later success in school (e.g., grade completion) and adult economic productivity (e.g., wages and employment).[13, 22]

1.2.2 Madagascar's Comprehensive Program

SEECALINE is a comprehensive community-level program, incorporating multiple activities that have been found to be associated with better child outcomes (see examples below). Community programs differ from individual treatment regimens in that they are typically made available to all (or most) residents of a community and sharing of information within a community is often encouraged. Examples of other community level ECD programs include national child health days,[23] breast-feeding promotion campaigns,[24, 25] provision of day care services,[26] conditional cash transfers,[27-29] and combinations of these.[28, 30] Community interventions have the potential to reach a large segment of the population.[30]

The Madagascar program includes growth-monitoring activities of children and nutrition education of their primary caregivers in a community center setting.[31] Trained nutrition workers weigh children (0 to 3 years of age) at monthly meetings, record the information in a growth chart, and provide the mothers with a private consultation if their children's growth falls below international guidelines for a given month. Children who are identified as severely malnourished (i.e., with a weight-for-height that is 3 standard deviations below the median of a reference population) are referred to a health clinic for treatment. The evidence of benefit to children from other large-scale growth monitoring and promotion programs is limited.[23, 32, 33] For example, improved growth was reported only among the youngest children in Uganda's Nutrition and Early Child Development Project, and no significant effects were reported for cognitive development of children.[15,33] This limited evidence of benefit is thought to be related to "coverage, intensity of contact, health worker performance, adequacy of resources, and the ability and motivation of families to follow advice."[33]

In addition to the growth monitoring activities, SEECALINE mothers are educated about the importance of exclusive breastfeeding during the first 6 months of a child's life, and of continued breastfeeding to 12-24 months. There is strong evidence of benefit from breastfeeding, specifically on child morbidity and mortality.[7, 25] However, there is inconclusive evidence of the effectiveness of large-scale breastfeeding promotion programs to improve growth in children.[25] In terms of cognitive benefit, most of the evidence is based on observational studies, as well as on the biological plausibility that a benefit would exist. For example, breastfeeding promotes intake of high quality nutrients and fatty acids that are essential for brain development.[34] In addition, breastfeeding benefits children by improving their immune response, limiting their exposure to pathogens from other food sources, and increasing maternal-child interaction.[35]

Perhaps the strongest evidence of a cognitive benefit comes from an experimental trial in Belarus that was modeled on the Baby-Friendly Hospital Initiative by the WHO and UNICEF. Investigators of the Belarus study reported higher means on all of the Wechsler Abbreviated Scales of Intelligence among breastfed children who were followed to 6.5 yrs.[36] However, no benefit of the Belarus program was found for height.[37]

Finally, SEECALINE participants are given guidelines for proper hygiene and complementary feeding practices, along with a cooking lesson and meal for attendees. In a systematic review of complementary feeding interventions in developing countries, parent educational programs were found to have a modest effect on child growth, with the most effective being those that emphasized feeding nutrient-rich animal-source foods.[38] None of the reviewed educational programs reported cognitive, language, or behavioral outcomes. Unlike some other nutrition-related programs (i.e., Ecuador's National Food Nutrition Program,[39] Bangladesh's Integrated Nutrition Project,[40] and Senegal's Community Nutrition Project),[41] SEECALINE does not provide fortified food or micronutrient, protein or energy supplements to participants.

1.2.3 Child Development Outcomes

Data collected over a 14-year period in Madagascar provides a rich source of information for testing the evidence of the program's effectiveness. The dataset includes a series of three nationally-representative cross-sectional anthropometric surveys that were administered in 1997/98, 2004 and 2011, with the first baseline survey administered prior to the beginning of the program roll-out. These surveys were administered in communities that became treatment and control sites after 1998. In addition to the repeated cross-sections, a prospective cohort of over 1000 children was followed from communities surveyed in 2004 when the children were less than 3 years of age. Seventy-five participating communities and 75 matched non-participating communities were randomly selected for this longitudinal follow-up.

A wide spectrum of developmental outcomes was assessed among children in Madagascar as part of the above surveys. Physical growth (height and weight) were measured in children under 5 years of age in each of the cross-sectional surveys, as well as in all households of children participating in the longitudinal cohort. Development across multiple domains (including physical, mental, and behavioral) was assessed in the longitudinal cohort at two follow-up periods, 2007 and 2011, when the children were 3 to 6 years and 7 to 10 years of age. In addition, mathematical and literacy achievement was evaluated in 2011. In my work presented here, for reasons described below, I focus on only two of these measures: weight-for-age and receptive language (i.e., comprehension of words, phrases and gestures used by others), as assessed by an adaptation of the Peabody Picture Vocabulary Test, version 3 (PPVT-III).[42]

I chose weight-for-age because it is one of the primary nutritional outcomes targeted for the program evaluation. Although prone to random error, I also chose weight because it is generally an outcome for which sources of systematic error have been previously identified and can be avoided during measurement. A prior analysis of the Madagascar program made use of the first two serial cross-sectional surveys to evaluate the impact on weight-for-age.[43] The authors of this previous analysis found that the program reduced the prevalence of child underweight in treated communities during a period of worsening malnutrition in non-treated communities (between 1997/98 and 2004).[43] I revisit these results in the context of a detailed framework for evaluation (see sections 1.4 and 1.5).

The reasons for my focus on vocabulary (a secondary, more distal outcome of the evaluation) are several. Language proficiency is an important domain of cognition and is integral to children's development across other domains of cognition (such as memory and problem solving), as well as to socio-emotional development. Children's ability to communicate their feelings and thoughts and to understand directions from parents and teachers is largely a function of their vocabulary and communications skills and is critical for their future success in school.[44] Early measures of vocabulary knowledge include lists of common words that caregivers check off as being understood (receptive) or spoken (expressive) by a child.[45] Expressive language (production of sounds, words, or phrases) can be measured as soon as children start to babble by counting the number and complexity of sounds an infant makes per minute. Similarly, receptive language can be assessed in infants by asking the mother what words a child understands or responds to. As children enter their preschool years (3 to 5 years of age), many tests of language, such as the PPVT, involve direct interaction with the child and no longer rely on the report of a caregiver. The PPVT can be used into adulthood and was used to assess the children's mothers in 2007 and the community nutrition workers in 2011.

Finally, early language ability is predictive of later school achievement across differing cultural contexts.[44, 46] Assessments of children's receptive vocabulary have been shown to be strongly correlated with poverty in other developing country settings,[47] as well as in Madagascar.[48] These results make it highly likely that my research has the potential for relevance outside of Madagascar. In the next section, I describe the importance of evaluating the performance of an instrument, such as the Peabody Picture Vocabulary Test (PPVT), after it has been translated and used in a context for which it was not originally developed.

1.3 Measurement – the case of language

In the U.S., there are well-recognized and accepted standards to promote the sound and ethical use of tests.[49] The stakes are high when results on a test can mean the difference between getting into a private school or college, or not; getting a job or a promotion, or not; receiving treatment, or not. As a result, great care is taken by U.S. test publishers to develop a "bias-free" test, avoiding bias that results in systematically lower or higher scores for a given sub-group of respondents (i.e., from differences due to gender, ethnicity, or socio-economic status). However, the development of a new test is a complex and time- consuming process and can be very expensive. For example, the multi-step process for developing the PPVT included extensive research on common English words used in the U.S., consultation with subject matter experts, iterative cycles of piloting items and performing reliability and validation checks, and establishing age and gender-appropriate standards using a representative sample of the U.S. population.[50] The PPVT was first created in 1959 and is already in its 4th revision; each revision attempting to reduce sources of bias (e.g., in each revision, images have been modified to be less culturally-biased).

In a series of articles on child development research in Africa, the authors repeatedly emphasize the need for "the design and local validation of developmental assessment tools."[51] However, due to budgetary and time constraints, U.S.-developed measures are commonly adapted and/or translated for use in developing countries (e.g., see the list of tests in the Nores & Bartlett review of early child development interventions).[52] The evaluation of the Madagascar program was no exception. Tests developed and carefully validated in one country are not guaranteed to maintain

their properties if they are adapted and translated for use in another country, and need to be re-validated in the new setting.[53] The process of translation and administration of the test in a different cultural and socio-economic setting may result in the loss of the test's psychometric properties and introduce bias into the study evaluation. For example, the difficulty of a test item may increase or decrease when translated, changing the overall order of item difficulty. A lack of knowledge of the age at which children achieve developmental milestones in the local context may result in some children hitting a stopping rule prematurely and being censored from taking later, potentially easier items.

These issues are discussed in a recent World Bank (WB) publication of a toolkit for adapting or developing instruments for assessing children in low-income settings.[44] While acknowledging the real world constraints, the authors of the toolkit offer practical advice for achieving test fairness that include (but are not limited to): careful selection of test materials that are culturally appropriate, working with local experts in adaptations and translation, and understanding how developmental milestones may differ in the local context. Many of these recommendations were followed in the Madagascar evaluation. The selection of the tests was based on their successful use in other developing countries.[48] For example, the PPVT has been translated into many other languages and has been used extensively throughout the world for assessing children's language skills. These uses of the PPVT include a multi-country prospective cohort study of the effects of poverty [54] and assessments of the effect of interventions, including nutrition-related experimental trials.[28, 52, 55] Prior to administration of the test in the Madagascar survey, the words for the PPVT were carefully translated in collaboration with a local clinical psychologist. Pilot runs were used to verify that Malagasy children understood and accepted the test. Although reliability checks were performed, no additional validation checks were completed prior to administering the survey, such as a comparison to a gold standard instrument (none exists for Madagascar).

In the absence of a gold standard instrument against which to validate the PPVT, it is of paramount importance that I establish the internal validity of the measure in Madagascar, after the fact, but before drawing inference about the program effect. In chapter 2, I explore three main areas of test reliability and validity for the PPVT. First, I assess the overall test and item level performance of the instrument using item response methods (IRM) separately from two survey years (when children were 3 to 6 years and 7 to 10 years of age). Second, I estimate the degree of differential item functioning by factors that have been frequently investigated in other settings with the same instrument (e.g., gender and language). Differential item functioning (DIF) is a form of bias that can be introduced if test takers with similar aptitudes from different subgroups give different responses to items on a test.[56, 57] Third, I evaluate the benefits of combining the data from the two survey periods in a two-dimensional model that allows for information to be shared across years. I describe the item response methods and discuss their relative strengths over classical test theory (CTT). Although I am unable to establish the external validity of the measure outside of Madagascar (i.e., how well the items and instrument would perform in another language and context), the results may be generalizable to Malagasy children in areas excluded from the study (i.e., provincial capitals). I present my lessons learned about working with the PPVT in Madagascar (e.g., how to test the instrument performance and administer the test in the field) and make recommendations for how these lessons can be applied in other developing country settings.

1.4 Identification of the target parameter – the case of pre/post data

In chapter 3, I apply the first steps of a roadmap for evaluating the effect of the Madagascar nutrition program on children's mean weight-for-age in a community.[58] These steps include specifying my research question and causal parameter of interest (the population average treatment effect or ATE), and assessing identifiability given the data that I actually observe. Importantly, I present the first part of the road map separately from the second part, which includes the methods of estimation and inference, and is presented in chapter 4. I emphasize this separation because it is common for the process of identification to be muddled with the process of estimation. In an introductory chapter on econometric evaluations of social programs, Nobel laureate James Heckman and co-author Edward Vytlacil point out that analysts often confuse the three main issues that face an evaluation: definition, identification and estimation. The authors state that "particular methods of estimation (e.g., matching or instrumental variable estimation) have become associated with 'causal inference' and even the definition of certain 'causal parameters' ..."[59]

As a consequence of this confusion, analysts faced with a new program evaluation may jump to a particular statistical method for estimating a causal effect. Investigators from different disciplines will bring distinct theoretical and analytical frameworks to the problem, which can lead to differing estimates of the causal effect and contradictory conclusions, in some cases without strong theoretical justification for the approaches used. The choice of method is particularly complex in observational studies, and made even more controversial with the availability of pre-treatment outcome data (as I will show in chapter 3). However, the selection of an estimator should happen after defining the research question and causal target parameter, and after the underlying assumptions necessary to identify the parameter are made explicit. By separating the evaluation of a program's effect into two chapters, I hope to underscore the need to "define first, identify second, and estimate last" (quoted from Judea Pearl's forward in *Targeted Learning* by van der Laan and Rose).[58]

For the Madagascar evaluation, I contrast defining the outcome as either the post treatment value or the change from pre to post treatment. Using these two outcomes, I identify three common statistical parameters that under different assumptions are equivalent to the causal target parameter of interest (the ATE) and I discuss the advantages and disadvantages of each.[60] I purposely include two difference-in-differences models that are popular in the field of econometrics for pre-post data (also known in social sciences as the change or gain score method).[60-63] I also include a common approach from the epidemiology literature in which the pre-intervention outcome (or lagged outcome) is included in the conditioning set of covariates. I avoid relying on parametric models whenever possible, using graphical models (directed acyclic graphs) to make the assumptions underlying the three causal models transparent. I demonstrate how the graphs can be used for locating sources of dependencies among variables. Finally, I highlight these dependencies with data simulations, and show how the estimate of the target causal parameter diverges from the truth when the necessary assumptions for a given model fail to hold. Although the context for this chapter is specific to the Madagascar study, the process is applicable to any program evaluation for which the investigator seeks to interpret an estimated effect of treatment on outcome as a causal effect.

1.5 Effect estimation – a comparison of methods

In chapter 4, I follow the second part of the roadmap for evaluating the Madagascar nutrition program.[58] The final steps of evaluation include estimation of my target parameter and inference (i.e., obtaining confidence intervals around the estimate). In chapter 3, I define my target parameter as the average treatment effect (ATE). In this next chapter, I evaluate whether the required assumptions for identifiability of the ATE are likely to hold in the context of the observed data from Madagascar, testing the assumptions where possible. Once identified, the causal target parameter can then be estimated, but a decision is needed about the choice of estimator, or method for estimation.

An estimator is a mapping function that takes as input an estimate of the distribution (e.g., my observed data as represented by the empirical distribution), and returns as output an estimate of the true target parameter value (e.g., the ATE). Ideally, the estimator is unbiased (the estimate is equal to the true target value), consistent (converges to the true value as the sample size increases) and efficient (has low variance of the estimate). A robust estimator is one where these desired properties are maintained under a wide set of conditions. The Madagascar study is essentially observational, as the program was assigned non-randomly and evaluated *ex-post facto*. In order to obtain an unbiased estimate of the program effect on mean weight-for-age, I require an estimator that ideally has the properties of a robust estimator described above. In other words, I want an estimator that does the best job possible of minimizing bias from observed confounders (due to covariate imbalance across treatment groups), as well as from the estimation process itself (i.e., does not introduce additional sources of bias). In chapter 4, I explore the characteristics and relative advantages of three candidate estimators: traditional parametric regression, inverse probability of treatment weights (IPTW or propensity score weighting), [64-66] and a relatively new method: targeted maximum likelihood estimation (TMLE).[58, 67]

Although commonly used, traditional regression has serious limitations for causal inference, such as the reliance on the correct specification of the parametric equation. IPTW addresses some of the issues associated with parametric regression, but is more sensitive to lack of experimentation within subsets of the covariates (i.e., if certain subgroups are never treated or always treated). TMLE is a doubly robust estimator with important advantages over the other two estimators. The method does not rely on parametric model assumptions, is expected to reduce the bias and variance of my effect estimate, and is flexible enough to explore target parameters either at the population level or at the individual level.[68] In addition, I present a non-parametric, data-adaptive approach for prediction (SuperLearner) to use in tandem with the IPTW and TMLE estimators. Finally, I compare estimates (and confidence intervals) for the ATE from these three different estimators, for each of the three statistical parameters described in chapter 3.

Once again, the context for this chapter is specific to the Madagascar study. However, the choice of estimator is relevant to any program evaluation for which the investigator seeks to obtain an unbiased causal effect estimate of treatment on outcome.

1.6 Implications

A careful evaluation of the impact of the Madagascar program on early child developmental outcomes is warranted. The results from my research may contribute to the gap in knowledge of

the effect of an at-scale nutrition education program on children's early development. As discussed previously, there is mixed evidence of the effectiveness of large-scale ECD programs. Although the lack of consistent results is generally attributed to possible problems of implementation and governance of the program, the failure to find a statistically significant effect (or alternatively, the success of finding one) may, in fact, be due to the types of problems described in my dissertation, or a combination of both (i.e., weak implementation and evaluation issues). There may be bias in the outcome (or other) measure, failure of the causal assumptions to hold, or bias from the method of effect estimation. Misleading estimates of a program's benefit (in either direction) have significant policy and funding implications in terms of making recommendations for the continuation, expansion, and/or replication of the program elsewhere. More importantly, the decisions we make based on the results of any ECD evaluation have important consequences for the children we are trying to help.

Chapter 2: Measurement – the Case of Language

2.1 Introduction

Tests developed and carefully validated in one country (such as the U.S.) are not guaranteed to maintain their properties if they are adapted and translated for use in another country. Once exported, the test needs to be re-validated in the new setting.[53] In this chapter, I focus on a translated and adapted version of the Peabody Picture Vocabulary Test, 3rd edition, version B (PPVT-III),[42] a measure of receptive vocabulary. I take advantage of existing data from an impact assessment of an early childhood nutrition program in Madagascar to explore three areas of reliability and internal validity for the PPVT. First, I assess the overall test and item level performance of the instrument using item response methods (IRM) separately from two survey years, administered to the same children, four years apart. Second, I estimate the degree of differential item functioning by two factors that have been frequently investigated in other settings with the same instrument: gender and language (or dialect). And third, I evaluate the benefits of combining the data from the two survey periods in a two-dimensional model that shares information across years. I present the item response methods that I will use for assessing these three areas, and discuss their relative strengths over classical methods.

2.1.1 Argument for IRT over Traditional

Classical test theory (CTT, also known as true score theory) is often used for evaluating test performance.[69] CTT assumes that each person has a "true score" for a particular ability (or characteristic) that would have been obtained if the ability had been measured without error. In other words, the observed score (typically the raw score across all items) for a test consists of a person's true score plus measurement error. This classical approach to evaluating ability is implemented primarily at the instrument level and fails to pick up some sources of bias that can occur at the item level.[56] Item response models (IRM), on the other hand, focus on modeling a subject's response to each item, allowing investigators to test for item-level bias.[70] The item response theory (IRT) underlying these models, also known as latent trait theory, is commonly used in education and psychology for the design, analysis, and scoring of multi-item tests or questionnaires that measure underlying (latent) abilities or attitudes. IRT is rarely used in public health. The simplest model in IRT is the unidimensional Rasch model where the probability of an observed response is modeled as a function of person (i.e., ability) and item (i.e., difficulty) parameters.[71] Specifically, in the Rasch model, the probability of a correct response is modeled as a logistic function of the difference between the person (θ) and item (δ) parameters:

$$P(X_{ni}=1|\theta_n,\delta_i)=f(\theta_n-\delta_i)=\frac{e^{(\theta_n-\delta_i)}}{1+e^{(\theta_n-\delta_i)}} \quad (1)$$

where X= the response (each item is scored as 0 or 1), n = n^{th} person, and i = i^{th} item. The mathematical unit of the Rasch model is the log-odds unit or logit and is applied to both the person ability and item difficulty parameters. When a person's ability is equal to the item difficulty (i.e., the logit difference = 0), there is a 0.5 probability of a correct response to that item for that person. As the person ability increases relative to the item difficulty, the probability of a correct response increases above 0.5. Similarly, as the person ability decreases relative to the item difficulty, the probability of a correct response decreases below 0.5. A logit difference of

+1 is equivalent to a probability of 0.73, and a logit difference of -1 is equivalent to a probability of 0.27.

By imposing the Rasch model, I accept that the relationship between the test item difficulties and the person abilities conforms to the statistical model shown in the above equation. To put it broadly, as the person ability goes up or the item difficulty goes down, the probability of a correct response goes up. In general, IRT is very useful for evaluating whether the items in an instrument are working well with the pattern of person responses obtained in a given sample. For instance, are respondents of similar ability mostly getting easy items right, hard items wrong and medium items right or wrong about 50% of the time? Information from an IRT model is useful for identifying poorly performing items (i.e., items that don't fit the Rasch model well) whose influence can be removed prior to estimating respondent's ability.

The common approach in settings where there are no local norms for scoring is either to calculate a raw score summary of item responses or to calculate age-standardized z-scores for the given analytic sample (the raw score is a strong function of age). Both approaches fail to take advantage of the information available at the item level and have the potential of introducing bias into the analysis, if the items did not retain their psychometric properties in the adaptation and administration of the test. Item response models (IRM) are probability models and go beyond raw scores: they rely on modeling responses to items by all subjects to estimate person ability and are not as sensitive to censoring or missing responses as a raw score total. The Rasch principle of specific objectivity states that as long as the Rasch model holds, then we do not need any particular set of persons to obtain the item difficulties, nor do we need to give every person the same set of items to obtain their relative proficiency estimates. IRT provides a clear benefit over the classical approach, especially in the case of missing response information.

2.1.2 Argument for additional DIF testing

One of the main threats to test validity comes from systematic differences (bias) when the test is administered to different groups within the same country or context. There are two ways that subgroup differences may occur. First, the subgroups may differ in their overall mean ability, commonly referred to as differential impact.[56] For example, children of highly educated mothers perform better on average on a test of intelligence than do children of uneducated mothers. Differential impact can be assessed with classical test methods and the subgroup can be treated as a confounder in impact evaluations (assuming the difference is not an effect of treatment). The difference may reflect a true difference in ability by subgroup and not necessarily a lack of fairness in the test.

The second subgroup difference is more subtle and referred to as differential item functioning (DIF). DIF is a measure of whether test takers from different groups (e.g., ethnic groups) with similar aptitudes give similar responses to items on a test. DIF is a measure of item-level bias where some items are easier or harder for one group than another. For example, suppose that more English language learners get a particular item wrong in an assessment of "speaking in English" than people who are native speakers, even when their overall ability is the same. This favoritism would constitute item bias by native language. The net effect of DIF may be that the groups differ in their overall mean ability if more items favor one group than the other. But this is not a requirement: DIF may exist when the mean ability is the same by subgroup (e.g., if the direction of DIF varies by item and is canceled out when the items are summarized into a total

score). As a result, DIF cannot be controlled like a confounder in a regression on the summary score.

Therefore, it is important to identify the presence/absence of DIF by key factors. The absence of differential item functioning (DIF) by group membership is an important property of a test and a key assumption in the item response model.[56] An advantage of IRMs (over the classical approach) is that they can be augmented to investigate DIF statistically by incorporating group membership into the model. The probability of a correct response becomes:

$$P(X_{ni}=1|\theta_n, g) = f(\theta_n - \delta_i + \gamma_i G) = \frac{e^{(\theta_n-\delta_i+\gamma_i G)}}{1+e^{(\theta_n-\delta_i+\gamma_i G)}} \quad (2)$$

where G is an indicator for group membership, and γ_i is an index parameter for DIF by G for item i.

The PPVT and DIF in the Madagascar Context

The PPVT consists of asking a subject to point to the correct image (out of a panel of four images) in response to a stimulus word spoken by the examiner. The images on a panel represent similar constructs or subject matter, for example, a panel might have an illustration of a candle, lantern, goose-neck lamp, and table lamp (four objects that provide light). The stimulus word might be candle and the distractors are the other three images. The distractor images are not intended to trick the respondent (e.g., they shouldn't sound alike) and should have an equal probability of being selected if the stimulus word is unknown.[42] Although originally developed in the U.S., the PPVT was chosen for Madagascar because a) it has been used throughout the world for intervention impact evaluations; [28, 52, 55] b) early language is predictive of later school achievement across differing cultural contexts; [44, 46] and c) assessments of children's receptive vocabulary have been shown to be strongly correlated with poverty in other developing country settings.[47] Although only available for purchase in American English or in Spanish (known as the TVIP), the instrument can be translated for use in non-U.S. cultural or linguistic settings upon permission of the publisher. Validation of the instrument in these settings is left to the investigator. However, the ability of investigators to address validity concerns may be limited by constraints imposed by the publisher.

Language background, gender and ethnicity are three common factors used to study DIF in language assessments.[72] The developers of the PPVT tested for item DIF by gender, race/ethnicity (i.e., White, African American, Hispanic, Native American, and other), and geographic region of the U.S., prior to finalizing the items included in the third version.[50] However, the PPVT has a history of controversy surrounding the performance of African American children on the test.[73] Researchers disagree as to whether the differential impact seen among African Americans is a true difference in ability or a function of item unfairness. It is likely that issues of poverty among African Americans are confounded with issues of ethnicity and culture. A similar problem may exist in Madagascar, where the dialect spoken is mixed with geography, ethnicity and socio-economic status.

The official languages of Madagascar are French and Malagasy. Official Malagasy is related to the Malayo-Polynesian languages of Indonesia, Malaysia, and the Philippines and is derived

primarily from Merina, a local Malagasy dialect. [74] Merina is spoken in the capital city and surrounding central highlands and has an ancient written tradition from which the official language was drawn. There are numerous other dialects spoken around the country that are region-specific and vary by ethnic descent. This variation is characterized by how different ethnic groups settled the island: Malayo-Polynesian groups settled in the central highlands, Arabic in the east and south-east, and African on the west coast. The dialects are mutually intelligible and have the same base syntax. For these reasons, Malagasy is considered a single language for translation purposes. The dialects differ in the pronunciation and vocabulary of certain words, with about a 70% similarity in lexicon.[75]

The extent to which the PPVT has been tested for language DIF in published results from other studies in non-English speaking countries is often unclear. The Young Lives study is an exception, as the investigators published a fairly detailed evaluation of the PPVT (or TVIP) in two age cohorts and four countries: India, Ethiopia, Vietnam, and Peru.[76] The authors of the Young Lives study report both classical test reliability information, as well as IRT reliability and validity results. Importantly, the reliability and validity information was assessed after "problem" items for a given country or age cohort were excluded. Problem items were items that failed to meet certain fit criteria and DIF by gender or language spoken (i.e., Spanish vs. Quechua in Peru). Based on the summary tables in their online annex, as many as 32 items of the first 72 PPVT items were excluded from the Ethiopian Amarigna version for the younger cohort (age 4.5-5.5 years).[77] A different set of items were deleted for the older Ethiopian cohort. Fewer items were deleted from the Indian and Vietnamese versions.

2.1.3 Argument for Multidimensional model

In a one-time administration of a test of intelligence or knowledge (such as the PPVT) participants are administered more than one item. The use of a multi-item instrument is equivalent conceptually to making repeated, independent measurements of the same construct (e.g., vocabulary knowledge) across many "mini" experiments. As the number of experiments (items) increases, the standard error of the estimate is reduced (assuming that the items actually represent the children's ability). Therefore, I gain precision in the ability estimates by analyzing responses to multiple items. With IRT, the theory relies on conditional independence of the items: for a given person proficiency, the items are assumed to be statistically independent of each other. Although it seems counter-intuitive for the items of a test to be uncorrelated, this assumption holds after conditioning on ability.[78] The assumption will fail if some dimension of ability that influences performance is not taken into account by the model. For example, if there is an unaccounted skill necessary for a correct response to an item, such as the ability to speak French, then the assumption fails.

I can improve the precision of the estimated person ability by repeating the administration of the test at multiple time points. However, I need to assume that the participants are not responding the second time by remembering their responses from the first time, in other words that their responses are independent between time points. One approach to utilizing repeated measures is to calculate a weighted mean of the estimates obtained for each person at each separate time point, where the weight might be the inverse of the absolute difference in ability over time. However, this approach does not take advantage of the correlation of the latent ability of each person between time points. An alternative approach is to use multidimensional IRT.

Multidimensional IRT models are an extension of unidimensional models, such as the Rasch model, and allow for subsets of items that measure different latent variables to be incorporated into a single model of a construct. With two administrations of the test, items administered in both years "anchor" the two periods together (assumes that their inherent difficulty has not changed) and allows for an estimate of the baseline skill and gain in ability over time. The theoretical framework that I use in this chapter is a latent growth item response model (LG-IRM) for two time points described by Wilson, Zheng, and McGuire. [79] Multidimensional item response models are considered more efficient than unidimensional models, providing better estimation accuracy of person abilities in one dimension (i.e., skill at time point 1) by using shared or collateral information from the correlated abilities in the other dimension (i.e., skill at time point 2). [80]

A Two-Dimensional Model in the Madagascar Context

I use a multidimensional item response model for the evaluation of the PPVT in Madagascar for the cohort of children who were assessed in 2007 and 2011. Although the experience of being tested in 2007 was certainly memorable for the children, I think that the first experience will not influence the responses in the second. First, the time between surveys was long (4 years) and the children were very young at the first administration (3 to 6 years). Second, the PPVT was one out of 9 tests administered over a 1 hour period in 2007, so it unlikely that the children remember the specifics of how they responded to any one test. Finally, the children were not given the correct answers, so I would not expect them to have learned new words in the PPVT from their first exposure. A direct comparison of the raw score totals between time points would not be informative since the test administration method differed by year. There was a high level of censoring in 2007 because the test was administered as recommended by the publisher (i.e., with stopping rules; see methods section for detail). In contrast, nearly all children were administered 72 items in 2011 (to avoid censoring) providing a much better estimate of vocabulary knowledge. I expect that the standard errors of ability estimates in 2007 will be reduced on average by "borrowing" information from these uncensored responses in 2011, an advantage that operates via the improved item estimates from the two-dimensional model.

2.1.4 Summary

In summary, I will use item response models to evaluate the performance of the PPVT in Madagascar, including: obtaining overall reliability estimates of the instrument, testing for individual item fit, and testing for differential item functioning. Based on this, I can recommend removing items with evidence of poor fit from the model, allowing items with differential item functioning to vary by subgroup, and using multidimensional IRM to gain shared information from the repeated measures.

2.2 Methods

2.2.1 Sample & Language Data

Children from 150 communities in all six of Madagascar's provinces were represented in the longitudinal cohort. Many of the same interviewers were hired in 2007 and 2011 to administer the PPVT. The test givers worked in pairs (one administering the test and one scoring), and received extensive classroom and field training prior to both surveys. All the words in the PPVT-III booklet B were translated, but only the first 96 words were ever administered to the children. In 20 of the items, no equivalent word exists in Malagasy. Therefore, these words

were “malagachisé” from the French equivalent (e.g., the French word for walrus was said with a Malagasy intonation). The Malagasy and French words were then back-translated into English by another party. Words without a Malagasy equivalent that were administered to the children are: kangaroo (15), ambulance (33), panda (38), dentist (43), hyena (75), walrus (80), and tropical (86). In addition, the French words for circle (10) and triangle (39) were used in 2007 because geometry is taught in French in Madagascar, and the Malagasy word is unfamiliar to children. However, these words were changed to the Malagasy equivalent in 2011. Finally, a few images were modified that were culturally inappropriate or ambiguous in the local language (i.e., a depiction of US dollars and cents was replaced with a picture of ariary and iraimbilanja, the currency used in Madagascar).

The PPVT is administered in sequential series of 12 items (panels of four images) and stopped if the respondent makes 8 or more mistakes in a series. Stopping rules prevent test fatigue, but rely on the series being ordered with increasing difficulty. In 2007, children were started at series 1 and continued through to series 6 or until they hit the publisher’s recommended stopping rule. Therefore, children in 2007 had a minimum of 12 items (from the first series) and a maximum of 72 items (all 6 series) administered. In an initial analysis of the 2007 Madagascar data, I found that the ordering of the items by difficulty had been changed by translation. Re-ordering of items is not allowed by the test developers and was not done. The risk of both right and left censoring is high from the use of start or stopping rules, if the order of item difficulties is gone. To avoid censoring of easy items occurring in high numbered series, children in 2011 (7-10 years of age) were started at series 3 and continued to series 8 without stopping, for a total of 72 items administered. A subset of 48 overlapping items was administered in both years.

A total of 1372 children took the PPVT in at least one of the years, 1244 children completed the test in 2007, 1224 children in 2011, and 1096 children in both years. Of these, 346 children had at least one common item administered both years. Note that in order to be included in this subsample, the children needed to have completed at least the first three series of items from the PPVT in 2007. This effectively restricted the subsample to a maximum of 388 children due to the use of stopping rules in 2007. The actual subsample of 346 is used for calibrating the common items that are used as anchors in the multidimensional model. How these children differ from the remainder of the sample is explored.

2.2.2 Classical Test Methods

Raw score totals were calculated from the items administered to each subject, with one point given for every correctly identified word. Respondents starting later than the first series are usually credited for any earlier series not administered, but this assumes that a) the order of item difficulty is not lost and b) the basal set (or the set of items that respondents can answer all or nearly all items correctly) is found for each child. In this analysis, the total raw score only includes items actually administered to a child. The published norm-referenced scores for the PPVT are based on a U.S. sample and are not an acceptable standard for children in Madagascar (they are not used here).[50]

Pair-wise Pearson or Spearman rank correlations were obtained between the scores and several demographic characteristics, including mother’s education and household wealth.[81] The household wealth index had been previously generated using principal component analysis to

aggregate wealth-related variables (i.e., asset ownership and dwelling characteristics such as electricity, running water, composition of floor, walls, and roof) into a single measure.[48]

Evidence of reliability, or evidence that the PPVT is yielding consistent results, was obtained in 2011 with the Cronbach's alpha indicator.[82] This indicator is a measure of internal consistency that identifies how well the responses on different items of the test are correlated. Excellent inter-rater reliability (>0.95) was obtained during the training session for the 2007 survey. Test-retest reliability data was not obtained.

2.2.3 Unidimensional Models

Separate unidimensional random coefficients multinomial logit (RCML) models were run on the data at both time points (n=1244 and n=1224 children in 2007 and 2011 respectively).[83] The responses to the items were dichotomized into correct/incorrect (1/0) score and a Rasch rating scale was used for the model. Item difficulty estimates were obtained based on the pattern of children's responses to item numbers 1 to 72 in 2007 and items 25 to 96 in 2011. Therefore, two sets of item difficulties were estimated for the 48 overlapping items 25 to 72, and estimates made for a total of 96 items. In each unidimensional model, the mean of the item difficulty parameters was constrained to zero by constraining the value of the 'last' item parameter to the negative sum of the other items.

In all IRT models discussed in this chapter, child ability estimates are constructed from their expected *a-posteriori* (EAP) distribution, where the latent ability distribution is assumed to be Gaussian (mean = 0, standard deviation = 1). EAP relies heavily on the distribution of the data, and therefore is sensitive to the sample population. For individuals with little data, the distribution is relied on more heavily, for those with a lot of data, less so. In addition, EAP underestimates the variance so that the EAP estimates are "shrunken," which would lead to bias in the item parameter estimates.[83] Alternative estimation procedures are available, such as maximum likelihood estimation (MLE), which is not influenced by the population model. However, EAP is the preferred technique in situations where there is a substantial amount of missing data.[83] Vocabulary items that were not administered to a given subject were treated as missing data. Therefore, EAP was chosen to handle the censored data in 2007, as well as the inherent missing data in the multidimensional model (to be discussed in the next section).

Wright maps are shown in the Appendix for the separate models. A Wright map is a useful visual representation of the estimation results as it places the item difficulties and person abilities on the same logit scale. On the left hand side there is a histogram of X's, illustrating the distribution of person ability, where each X represents a subset of respondents (noted at the bottom of the map). The item responses are located on the right side of the map at the point where a respondent has a 50% chance of responding correctly to the item (also known as the Thurstonian threshold).[84] Persons with abilities above the threshold have a greater than 50% chance of getting the item right and persons below the threshold have less than a 50% chance. A quick glance at the Wright map can tell me a) the approximate item ordering from easy to hard, b) whether the distribution of the person abilities is approximately normally distributed, and c) how well the item difficulties cover the full range of person ability.

Standard errors of measurement (SEM) are obtained for both the item difficulty and person ability parameter estimates. The SEMs for item difficulty are generally small in large samples,

since there are a large number of responses to any one item. Similarly, the SEMs for person ability are a function of the number of items administered to the respondent. When plotted against the ability estimates, the SEMs for ability will typically take on a U-shaped pattern if there is a good overlap between item difficulty and person ability. Specifically, the errors are lowest for respondents who were administered a large number of items centered on their ability. The errors increase for respondents with ability estimates at the extremes where there are fewer items with difficulties close to their location.

For evaluating the fit of the items to the model, I use a weighted mean square (MNSQ) fit statistic, or infit, which is a measure of how well the Rasch model fits the observed responses for an item.[84] Specifically, the infit statistic is a ratio of the variances of the observed residuals over expected residuals. Therefore, an infit equal to 1 indicates that the observed residuals vary as much as would be expected by chance. Infit values above 1 denote positive misfit, or more variation than expected. Infit values of less than 1 denote negative misfit, or less variation than expected. Deviations from 1 are evidence of lack of fit, but some deviation is expected due to random measurement error. The infit is considered by other researchers to be within an acceptable range if it falls between 0.75 and 1.33 (=3/4 and 4/3).[84] Evidence of statistically significant misfit is obtained if a t-statistic is greater than +/- 2 (the t-statistic is based on a transformation of the infit into a standard normal distribution).[84] Negative misfit is usually less of a concern than positive misfit, as negative misfit generally represents a highly discriminating item. However, having multiple items that are highly discriminating is redundant and may extend the length of the test unnecessarily. Although it would not be surprising to find some significant misfit given my large sample size, I will look for patterns among the items that demonstrate significant misfit.

Item characteristic curves provide a graphical representation of the fit of the item. The probability of a correct response is plotted as a function of person ability and has an S-shape. Persons with low ability with respect to the item difficulty are expected to have a low probability of success on the item, whereas persons with high ability with respect to the item have a high probability of success. The plots included in this chapter show both the empirical item characteristic curve (based on the observed data) and the modeled curve.

Finally, a second measure of instrument reliability is available from the item response model, which should give comparable results to the Cronbach's alpha indicator. The person separation reliability indicator is calculated as the difference in the observed total variance of the estimated abilities and the residual variance not explained by the model, divided by the variance explained by the model.

Differential Item Function by Subgroup (Unidimensional)

Differential impact and differential item functioning was assessed for gender and language spoken in the home in the unidimensional models. A dichotomous indicator variable was used for language based on the following survey question posed in 2011 to the primary caregiver for the child: "What language do you speak with your child at home: official Malagasy, French, or a local dialect?" Missing data from children only observed in 2007 was imputed from the community median response. Since French was selected only a couple of times, these observations were replaced with the community median as well. Although, testing for

differences by province or geographic location would capture more detailed differences amongst the dialects, this analysis was not performed.

Differential impact and DIF are tested by adding two terms into the unidimensional Rasch model: a term for group membership (e.g., gender) and an interaction term between the item and the group. The parameter estimate obtained for the group membership is an estimate of the overall mean difference in abilities between the subgroups (e.g., the difference between boys and girls). If this parameter estimate is more than 1.96 times its standard error, the difference is considered statistically significant.

A single parameter is estimated per item*group for one of the subgroups (e.g., the girls). The parameter for the other group is constrained as the negative of this estimate (centered on zero). The DIF effect size is then calculated as two times this item*group-specific parameter. The effect size is categorized as negligible (<0.426 logits), medium (≥0.426 and ≤ 0.638 logits), or large (>0.638 logits). This classification scheme is based on a log transformation of the Mantel–Haenszel (MH) common odds ratio and is consistent with the Educational Testing Service (ETS) DIF categories. [85] DIF of an item was considered statistically significant if the effect size was greater than 1.96 times the standard error of the item*group parameter estimate. Items were examined if there was evidence that the DIF was both statistically significant and if the effect size of the DIF was larger than 0.638 logits.

Throughout the chapter, a reference to an item exhibiting DIF implies that the DIF was large and statistically significant, unless otherwise specified. Some items may have small to moderate DIF that is significant or large DIF that is not significant, but these items are not typically included in the DIF item counts discussed in this chapter.

2.2.4 Multidimensional Model

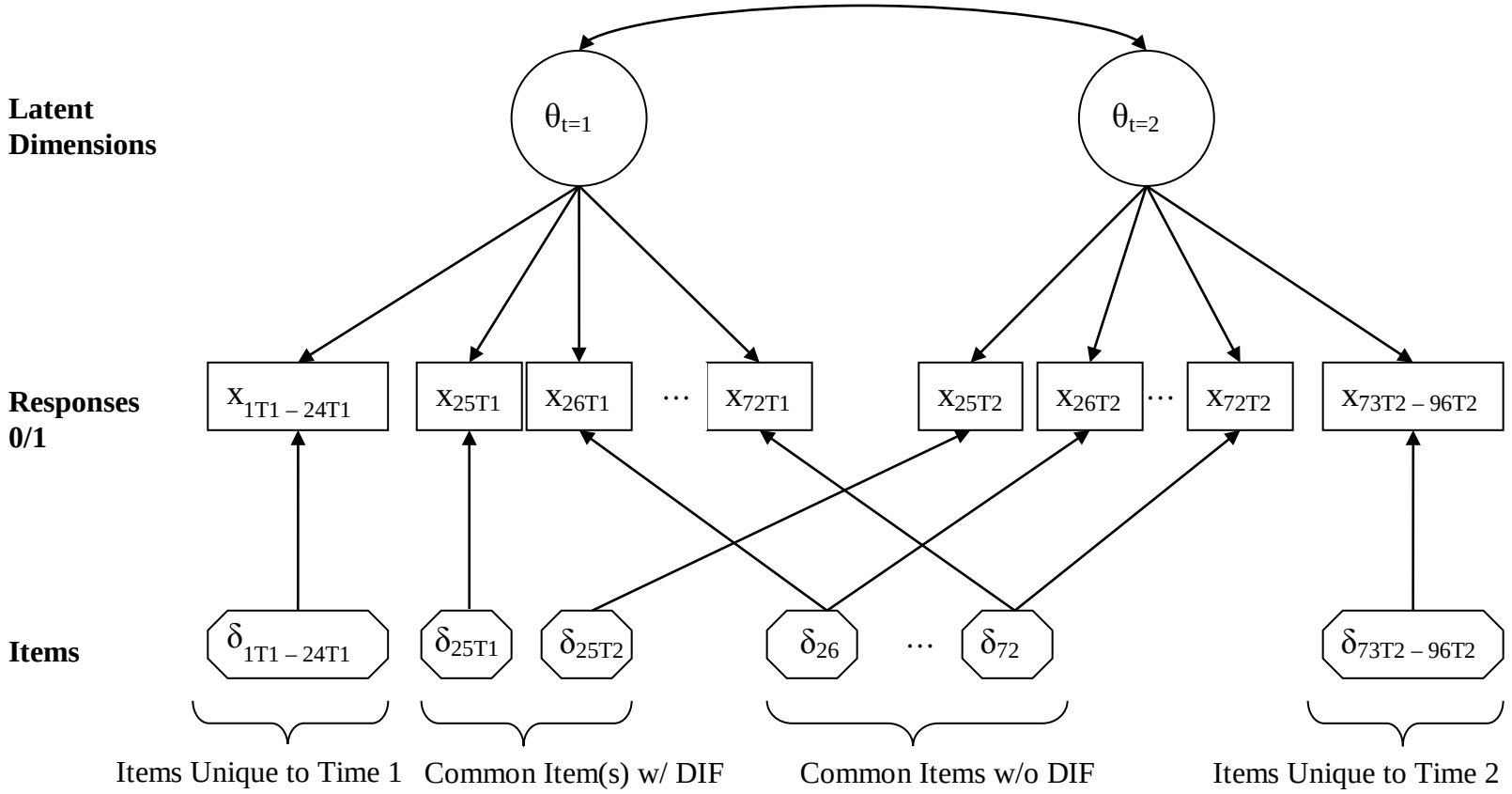
Many of the same methods discussed in the above two sections apply to the multidimensional model. However, building a multidimensional model is more complex than implementing the unidimensional models. Figure 2.1 depicts the relationship between the latent language abilities from the two time periods, the observed responses to individual vocabulary items, and the estimated difficulties of those items. The details of building the model are described next.

Anchor Item Identification & Calibration

The first part of the process involves identifying the set of items administered in both years that will act as anchors in the multidimensional model. Anchored items are required to have constrained difficulties that do not differ from one year to the next. In other words, the anchors cannot exhibit DIF by the year that they were administered. In this way, children can be assessed on the same quantitative scale in both years. The estimation of these anchored item difficulties are obtained from the restricted subset of children who were administered at least one item in both years from the 48 common items (n=346).

First, separate unidimensional calibrations were repeated on the sub-sample to see if the overall item fit statistics differed from those obtained from running separate models on the full sample. For a simple evaluation of DIF, the item difficulty estimates in 2007 were plotted against the estimates from the 2011 data for the overlapping 48 items. Items were flagged that differed by

Figure 2.1: Theoretical IRT-Based Model for Measuring Language Acquisition



more than 0.638 logits between years, and compared with the concurrent differential item functioning analysis described in the next paragraph.

The set of anchor items (out of the 48 common items) was identified in a more formal, iterative process. Specifically, DIF by year was tested in a concurrent analysis of the responses to the common items in both years using a unidimensional model (one row per person per year). An interaction term for the item times year was added to the model. Items were removed as anchor candidates if there was evidence that the DIF was both statistically significant and if the effect size of the DIF was larger than 0.638 logits (see method for DIF testing by subgroup in section 2.4).[85] The process of testing for DIF was repeated with the remaining items, until no additional DIF was found in a final subset of common items.

Once the anchor items were identified, a two-dimensional item response model was run on the responses to the anchor items in both years from the subset of 346 children. This constitutes the calibration step that generates a set of item difficulty parameters that will be constrained in the full model (see next section). The parameters generated in this calibration step were exported, with the parameter for the last item computed as the negative sum of other items.

Complete Multidimensional Model

The second part of the model building process involves constraining the parameters for the anchor items, obtaining item difficulty estimates for the remaining items, and estimating person ability in both years. This is done by importing the anchor item parameters (as described above) into a new two-dimensional model that now includes all of the items and responses for all of the children. The non-anchored item difficulties are estimated by this new model and allowed to vary in difficulty by year. These items include items administered in only one year and common items that failed the DIF test by year. In this way, the maximum number of items is retained and can contribute to the ability estimates of the subjects. Because the anchored items have constrained difficulty parameters, no additional constraints were set.

As with the unidimensional models, the Wright maps, standard errors of measurement and item fit are inspected. The multidimensional Wright map shows the ability estimate by year with two histograms and the overall difference in the mean child ability by year is calculated. Items administered in only one year and the items with DIF do not contribute to the "gain" estimate between years, but they are contributing to the accuracy of the subjects' ability estimates. Correlations of the two dimensions are obtained directly from the software package and are dis-attenuated for measurement error.

Differential Item Function by Subgroups (Multidimensional)

Differential item functioning by gender and language was repeated with the two-dimensional model. The method differs from that described in section 2.2.3 in two ways. First, all the item parameters from the complete two-dimensional model are exported for the 96 items (anchor and non-anchored items). Next, separate unidimensional models are run for each year, importing and constraining the item parameters from the two-dimensional model (items 1-72 for 2007 and 25-96 for 2011). As before, an item*group interaction term is added to the model to test for DIF. Since the item difficulties are all constrained, no additional constraints are used.

Unlike previously, the item*group parameters are not constrained to be centered at zero in this model. Therefore, two parameters are estimated for each subgroup. The DIF effect size is calculated as the difference between these two item*group parameters. The same DIF effect size categorization and significance testing was performed as before. Note that the standard error of the DIF effect size was roughly calculated by taking the square root of the average variance of the two estimates.

2.2.5 Software

ACER ConQuest version 2.0 was used for all of the IRT modeling.[83] Stata/MP 10.1 for Windows was used for cleaning and generating the dataset, and for raw score statistics.

2.3. Results

2.3.1 Sample Characteristics

Just over half of the children in the sample (assessed in either year) are female (51.4%). The mean age was 54.6 months (SD 10.4, range 33-73) and 103.2 months (SD 10.5, range 81-126) in 2007 and 2011, respectively (see table 2.1, last column). Less than a fifth of the sample reside in urban areas (17.9%) and they are fairly evenly distributed across the six provinces in Madagascar. The fewest reside in the remote northern province of Antsiranana (6.1%) and the largest percentage in the southern province of Toliary (27.4%). Approximately 24% of the children's mothers' are uneducated, about half have some primary school education (55.6%), and fewer than 21% achieved secondary or above education.

A dichotomous indicator for language was set equal to 0 if the primary caregiver reported that the language spoken at home with the child is official Malagasy, or equal to 1 if a local Malagasy dialect other than the official language is spoken. Over three quarters of the children speak a local dialect in their home (76.4%). Language spoken is almost entirely a function of province with approximately 61% of official Malagasy speakers living in the central province of Antananarivo (where the capital is) and another 23% in the neighboring province of Fianarantsoa. Language spoken is also a function of household socio-economic factors. Among the mothers with no education, the percentage of children who speak a local dialect at home goes up to nearly 92%. And among the poorest households (the bottom wealth quintile in the sample), the percentage of children who speak a local dialect is 95% versus 58% in the wealthiest households (top wealth quintile).

Nearly 12% of children were reported by their mothers to have some type of developmental delay, or hearing or sight disability. Note that these reports may not be based on a doctor's diagnosis and are subject to inaccurate reporting. Specifically, two questions were asked of the primary caregiver: 1) Does your child have hearing problems or difficulty in seeing, either during the day or at night; and 2) Compared with other children his/her age, do you think that your child has a serious delay in his/her mental development? The reports of delay or disability did not differ significantly by household wealth or mother's education.

The subset of 346 children with at least one common item from both years were on average 4-5 months older, more likely to speak official Malagasy at home, to live in an urban location or in the capital province, and to have a mother with secondary or above education. These results are consistent with children who are more likely to have completed at least 3 series of items in 2007

before being censored. Even so, all demographic groups are reasonably well represented by the subgroup (see table 2.1, third column). These results, combined with the IRT assumption of specific objectivity (we do not need any particular set of persons to obtain the item difficulties), suggest that the subset group is adequate for obtaining item difficulty estimates on the anchor items for the two-dimensional model.

Table 2.1: Summary demographics by inclusion in the anchor item estimation

	No common items	**Had common item(s)**	**All children**
N	1026	346	1372
Female (%)	51.4	50.3	51.1
Mean age in '07 (range)	53.2 (33-72)	58.6 (36-76)	54.6 (33-76)
Mean age '11 (range)	101.6 (81-126)	107.1 (83-124)	103.1 (81-126)
Speaks local dialect† (%)	80.8	63.3	76.4
Urban location (%)	15.0	26.6	17.9
Concern for developmental delay or impairment† (%)	11.2	13.0	11.7
Wealth Quintiles (%)			
1st	21.7	16.5	20.4
2nd	22.0	12.7	19.7
3rd	22.1	14.2	20.1
4th	19.2	24.0	20.4
5th	13.7	32.7	18.5
missing	1.2	0.0	0.9
Mother's education (%)			
none	26.1	16.2	23.6
primary	56.5	52.9	55.6
secondary or above	17.4	30.9	20.8
Province(%)			
Antananarivo	13.1	21.7	15.2
Fianarantsoa	21.9	21.4	21.8
Toamasina	19.1	13.3	17.6
Mahajanga	12.3	10.7	11.9
Toliary	26.6	29.8	27.4
Antsiranana	7.0	3.2	6.1

† The variable for local dialect is a binary indicator that the child speaks a Malagasy dialect other than the official Malagasy language at home. The variable for concern is a binary indicator for a yes response to either of two questions about hearing or sight impairment, or mental development as compared to other children of the same age.

Classical Test Scores

In 2007, the mean raw score for children is 13.6 words, but the distribution is right-skewed with a median of 11 words, and range of 1 to 57 words (SD 8.7 words) (see table 2.2). Approximately 10% of children were administered only 12 items, stopping after the first series, and another 54%

were stopped at the second series. Less than 5% of children made it to series 6. The raw score total distribution is bimodal, with a peak around 4 words for children who were censored at the first series, and a second peak at 10 for those who made it to the second series or beyond (see Appendix A1, figure 2.7 for histogram). In 2011, nearly all children were administered 72 items since the stopping rule was ignored. The mean raw score total in 2011 is 30.5 words, with a more normal distribution than in 2007 (median 29, range 11 - 60, and SD 9.5 words). As expected, the raw score is strongly related to age, with an average gain of about 1 word for every 4 months of increasing age in both periods. The average gain in ability from 2007 to 2011 cannot be directly estimated as the scores are on two different scales (and credit was not given for items that were not administered). However, the raw scores in 2007 are positively correlated with those from 2011 with a Pearson correlation of 0.42 (table 2.2).

Table 2.2: Comparison of Models

	Raw Scores	**2 Separate Unidimensional**	**Multi-dimensional**
Items used to estimate ability	12 – 72 items	72 items	96 items
Model fit	N/A	142845	142939
Mean (SD) 2007 ability	13.6 (8.7) words	-0.86 (0.49) logits	-1.07 (0.47) logits
Mean (SD) 2011 ability	30.5 (9.5) words	-0.37 (0.57) logits	-0.12 (0.57) logits
Mean "gain" in ability	N/A	N/A	0.95 logits
Correlation of '07 and '11 scores[†]	0.424	0.437 (dis-attenuated: 0.598)	0.588
Median (range) 2007 SE	N/A	0.37 (0.21-0.49)	0.34 (0.15-0.44)[‡]
Median (range) 2011 SE	N/A	0.25 (0.24-0.30)	0.25 (0.14-0.28)[‡]
Separation Reliability 2007	N/A	0.641	0.636
Separation Reliability 2011	N/A	0.832	0.777

[†] Raw and unidimensional score correlations are Pearson correlations. The raw score correlation is not corrected for measurement error. Multidimensional score correlation is obtained directly from Conquest and are corrected for measurement error

[‡] Excludes standard error of estimates for children not administered the test in a given year

The raw scores were significantly correlated with mother's education (coded as none (0), primary (1), or secondary and above (2)) and household wealth index, with a larger positive correlation for the older cohort (see table 2.3). The Pearson correlation of the score and household wealth index is 0.35 and 0.53 in 2007 and 2011, respectively. Speaking a Malagasy dialect at home other than official Malagasy (using the binary indicator described previously) is negatively correlated with the score (Spearman rank correlation of -0.23 and -0.41 in 2007 and 2011, respectively). Neither gender nor report of developmental delay/disability was significantly correlated with the scores in either year.

2.3.2 Separate Unidimensional Models

Separate unidimensional IRT models were run on all participating children in 2007 and 2011.

Instrument Reliability

The Cronbach's alpha in 2007 is not reported here due to the high percentage of response data that is missing for the administered items (59%). The person separation reliability is moderate at

0.64, which is in part due to the relatively small number of items administered to children who were stopped. The low reliability can also be explained by the poor overlap between the item difficulties and the children's estimated abilities (see Wright Map in Appendix A1 figure 2.9). In other words, the first 72 items of the PPVT were too hard for most of the children.

In 2011, the alpha coefficient is very good at 0.85 and the person separation reliability is reassuringly comparable at 0.83. Most of the children in 2011 were administered all 72 items, with less than 1% of the response data missing. In addition, there is a good overlap between item difficulties and person abilities in 2011 (see Wright Map in Appendix A1 figure 2.10).

Table 2.3: Correlations of test scores with demographics

	2007			**2011**		
	Raw score	**1D IRT score**	**2D IRT score**	**Raw score**	**1D IRT score**	**2D IRT score**
1D IRT score	0.89	1		0.999	1	
2D IRT score	0.92	0.97	1	0.995	0.996	1
Female	-0.02	-0.02	0.00	-0.05	-0.05	-0.02
Speaks local dialect†	-0.23	-0.24	-0.32	-0.41	-0.41	-0.40
Urban location	0.20	0.20	0.19	0.16	0.16	0.15
Concern for developmental delay or impairment†	0.02	0.03	0.04	0.02	0.02	0.03
Maternal education†	0.21	0.21	0.29	0.39	0.39	0.38
Household wealth (index)	0.35	0.32	0.41	0.53	0.53	0.52

† The variable for local dialect is a binary indicator that the child speaks a Malagasy dialect other than the official Malagasy language at home. The variable for concern is a binary indicator for a yes response to either of two questions about hearing or sight impairment, or mental development as compared to other children of the same age. Mother's education was coded as ordinal categories: none (0), primary (1), or secondary and above (2).

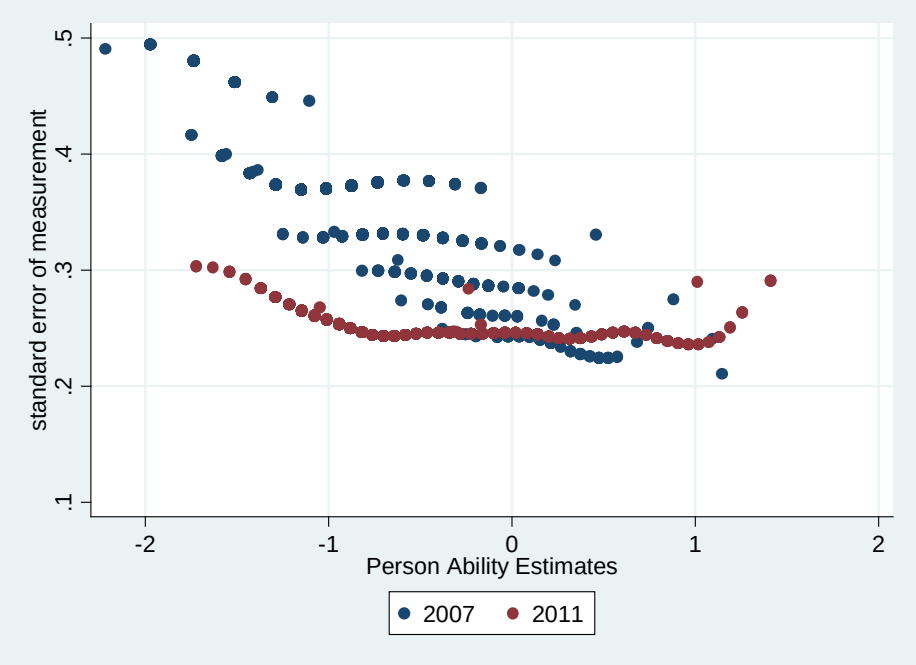
Respondent Measures

The ability estimates obtained from the separate IRT models range from -2.2 to 1.1 logits in 2007 to -1.7 to 1.4 logits in 2011. The distribution of the IRT ability estimates in 2007 has a much more normal appearance than the raw scores. However, the IRT estimates are highly correlated with the raw totals, as expected (see table 2.3). As with the raw scores, the average gain in ability from 2007 to 2011 cannot be estimated as the estimates are on two different scales. The ability estimates in 2007 are positively correlated with those from 2011 with a Pearson correlation of 0.44. Correcting for attenuation by measurement error gives an estimated correlation of 0.60 (dis-attenuated by dividing the correlation by the square root of the product of the reliability coefficients of the two years).[86, 87]

The standard errors of measurement as a function of ability estimates are shown in figure 2.2 for both 2007 and 2011. The slightly U-shaped pattern obtained in 2011 is typical for item response models. The error increases for respondents with ability estimates at the extremes where there

are fewer items with difficulties close to their location. This also explains the distinct pattern in 2007 for the number of series administered to the children, with increasing standard errors as the amount of censoring increases. For the approximately 10% of the children administered a single series (5 series censored) the standard errors ranged from about 0.45 to 0.5 logits, and the 49% administered 2 series had standard errors approximately 3.5 to 0.4 logits. Two children answered fewer than 12 items in 2007, had very high standard errors, and were dropped from the analysis. The children administered all 6 series (72 items), in either year, have the lowest standard errors between 0.2 and 0.3 logits.

Figure 2.2: Standard error of measurement for separate unidimensional models by year



A respondent with a standard error of measurement of 0.25 logits has a 95% confidence interval (CI) of ± 1.96*0.25 logits (approximately1 logit) around their estimated ability (e.g., for an estimated ability of 0 logit, the CI is -0.49 to 0.49 logits). This 95% CI is about a third of the full range of the respondent locations in 2007 and a fifth of the range in 2011. Children who were censored after the first series have confidence intervals that are nearly double that at 2 logits wide.

Internal Structure

The internal structure of the instrument was checked at the instrument level and at the item level with the use of item fit statistics and Wright maps from the two separate unidimensional Rasch models.

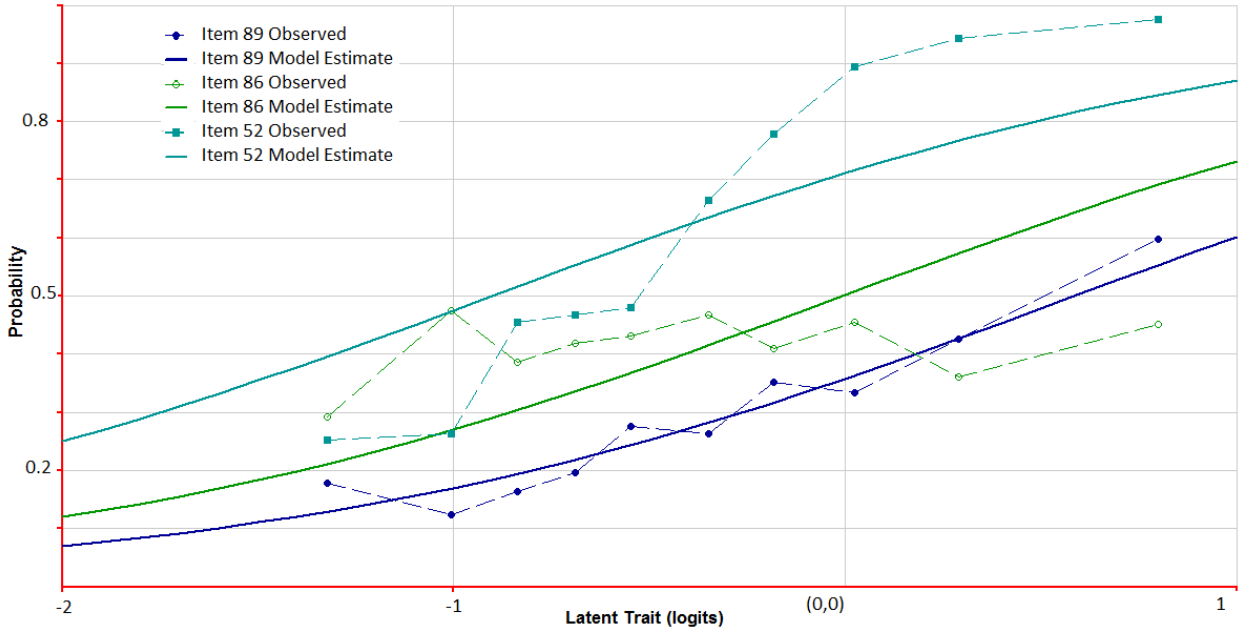
The table of item statistics and the Wright map for 2007 show that the item difficulty estimates are distributed from approximately -4 to 2 logits with most items clustered to the high end (see Appendix A1 table 2.4). The respondents have a skewed distribution of abilities centered well below zero, near -0.9 logits. A comparison of the distribution of item difficulties against the distribution of person abilities indicates that the items were more difficult overall for the

respondents than desired in 2007. In 2011, the difficulty estimates are reasonably distributed from approximately -2.7 to 2.1 logits. The respondents have a similar distribution of abilities centered near zero. The item difficulties are dispersed fairly evenly across the full range of person abilities in 2011.

The order of item difficulties is described by the publisher as a) within a series of 12 items, the easiest 3 items are given first and the hardest 3 items are given last, and b) items in a series increase in overall difficulty as the series number increases (i.e., all items in series 1 should be easier than all items in series 2).[50] The item difficulty estimates obtained in Madagascar indicate that the ordering was lost in both years, as was expected. For example, item 68 in series 7 (tortoise) was easier than almost all of the other items, whereas item 29 from series 3 (coin) was harder than most.

The item difficulty estimates were inspected for all items administered in both years. In 2007, the standard error of measurement for the item difficulties ranged from 0.06 to 0.25 logits, with the exception of item 72 (SE 1.2 logits). The estimate for item 72 was constrained to be the negative sum of the other difficulty estimates and was administered to 59 children in 2007 (<5% of the sample). In 2011, the standard errors on the item estimates were much smaller on average, ranging from 0.05 to 0.06 logits, with the exception of the last constrained item 96 (SE 0.43 logits). These results are consistent with the larger number of responses obtained in 2011 for all of the items.

Figure 2.3: Three item characteristic curves by scores in 2011 (items 52, 86 & 89)



Overall, the fit of the item responses to the unidimensional Rasch model was reasonable. The weighted mean square fit statistics (infit) for all of the items are within the acceptable boundaries of 1.33 and 0.75 (see figure 2.12 in Appendix A1). Although the items were within the acceptable bounds, the infit differed significantly from 1 for a number of items (t-statistic for the infit less than -2 or greater than 2). Given the large number of items tested for misfit and the

large sample size, I would expect to find some statistical evidence of misfit. However, the t-statistic for the infit was as large as -10 (item has less variation) and +8 (item has more variation). In 2011, 18 items exhibited statistically significant negative infit and 13 items had statistically significant positive infit. Examples of three item characteristic curves are shown in figure 2.3: items 52, 86, and 89. Item 89 (river) is an example of an item with excellent fit (infit 0.99, t-statistic -0.3). Item 52 (huge) is an example of an item with negative infit (infit 0.85, t-statistic -7.8). And item 86 (tropical) is an example of an item with positive infit (infit 1.16, t-statistic 8.2). Notice that the observed probabilities are fairly flat across the range of person ability for item 86 (less discriminating), whereas item 52 has a steeper curve than expected by the model (more discriminating).

I looked for patterns that might explain the evidence of statistical misfit. First, I checked for patterns by type or difficulty of the words. Over a third of the items with significant negative misfit were action verbs, with item 30 (peeking) having the strongest evidence in both years. However, nearly a third of the items with significant negative misfit were also action verbs. Perhaps more interestingly, half of all of the French words administered to the children had significant positive infit in one or the other year of test administration.

Next, I checked for patterns that might be associated with the length of the test and choice of distractor items. The local Malagasy clinical psychologist had noted that children who were bored towards the end of the test tended to pick either the top right (#2) or bottom right (#4) image in the panel, sometimes even before the word was spoken by the test giver. Her observation is supported by the data. The distractors were clearly not chosen at random. In exactly half of the items, the most common distractor image selected was #2 (upper right). The next most common distractor was image #4 (bottom right). The bottom left image (#3) was the least likely distractor to be selected (only 8% of the time). Therefore, I can assume that more children will get an item right than expected by chance alone if the correct response corresponds to image #2. In fact, nearly half of the items with significant positive infit were items with image #2 as the correct response in 2011. It appears that if the children are tired from the length of the test and don't know the word, they are more likely to pick image #2. This is reflected by the positive misfit items including more, higher numbered items (~60% were from the last 2 series in 2011). Test fatigue probably would have been alleviated if the items had been re-ordered according to the local difficulty (subject to obtaining permission by the publisher).

DIF by gender and language

Differential item functioning was tested for gender and language spoken in the home.

Gender

In 2007, the estimated difference in overall ability (impact difference) by gender is 0.11 logits (SE 0.012), with girls finding items harder on average than boys. The actual parameter estimate for girls is 5 times larger than its standard error, so the difference by gender is significant. In 2011, the overall difference by gender is much smaller, 0.05 logits (SE 0.008), but still statistically significant and favoring boys.

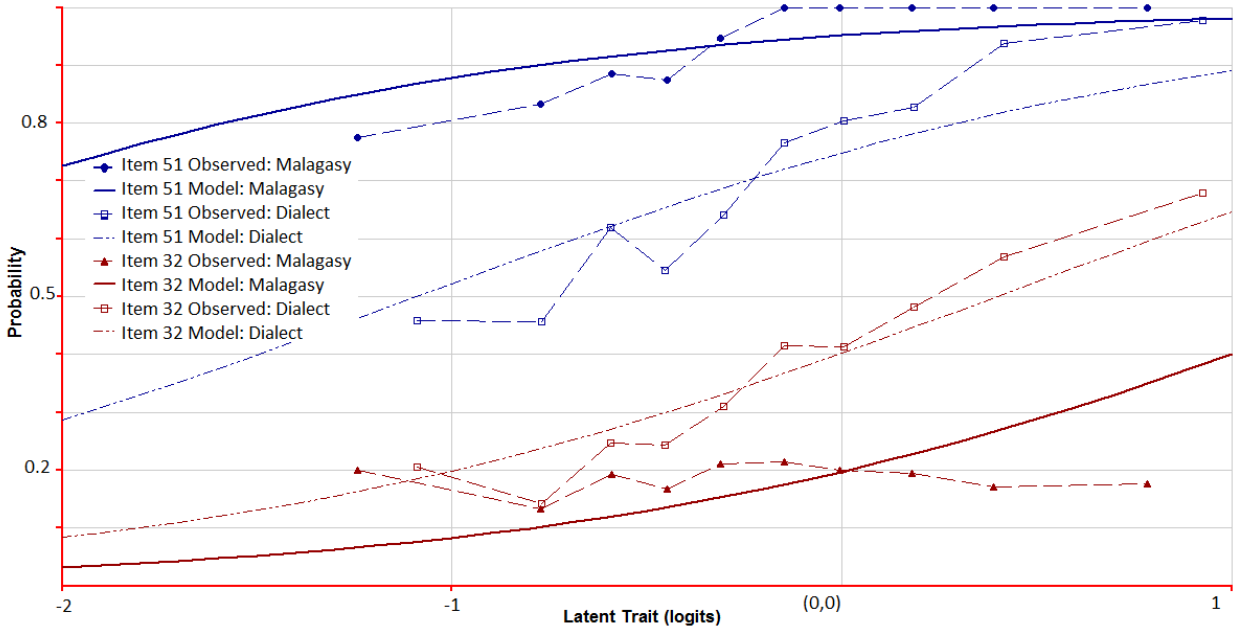
At the item level, 5 items demonstrated large and statistically significant DIF by gender in 2007, and none in 2011. In 2007, two of the items favored girls (items 26 and 60) and three items

favored boys (51, 54, and 64). Item 72 has a gender DIF effect that also favors boys of nearly 1 logit, but was not statistically significant in this model, most likely due to the small number of children who completed the item. There were no obvious patterns found that explained the evidence of gender DIF.

Official Malagasy vs. Local Dialect

In 2007, the estimated difference in overall ability (impact difference) by language is 0.27 logits (SE 0.012), with children who speak a local dialect other than official Malagasy scoring lower (on average) than those who speak official Malagasy. The DIF effect estimate for speaking a local dialect is ten times larger than its standard error, so the difference by language is significant. In 2011, the overall difference by language is more than double, 0.68 logits (SE 0.008), and also statistically significant. Again, children who speak a local dialect scored lower on average. These impact estimates are equivalent to about 10% in 2007 and 20% in 2011 of the full range of the ability estimates.

Figure 2.4: Item characteristic curves for two items (32 & 51) with language DIF in 2011

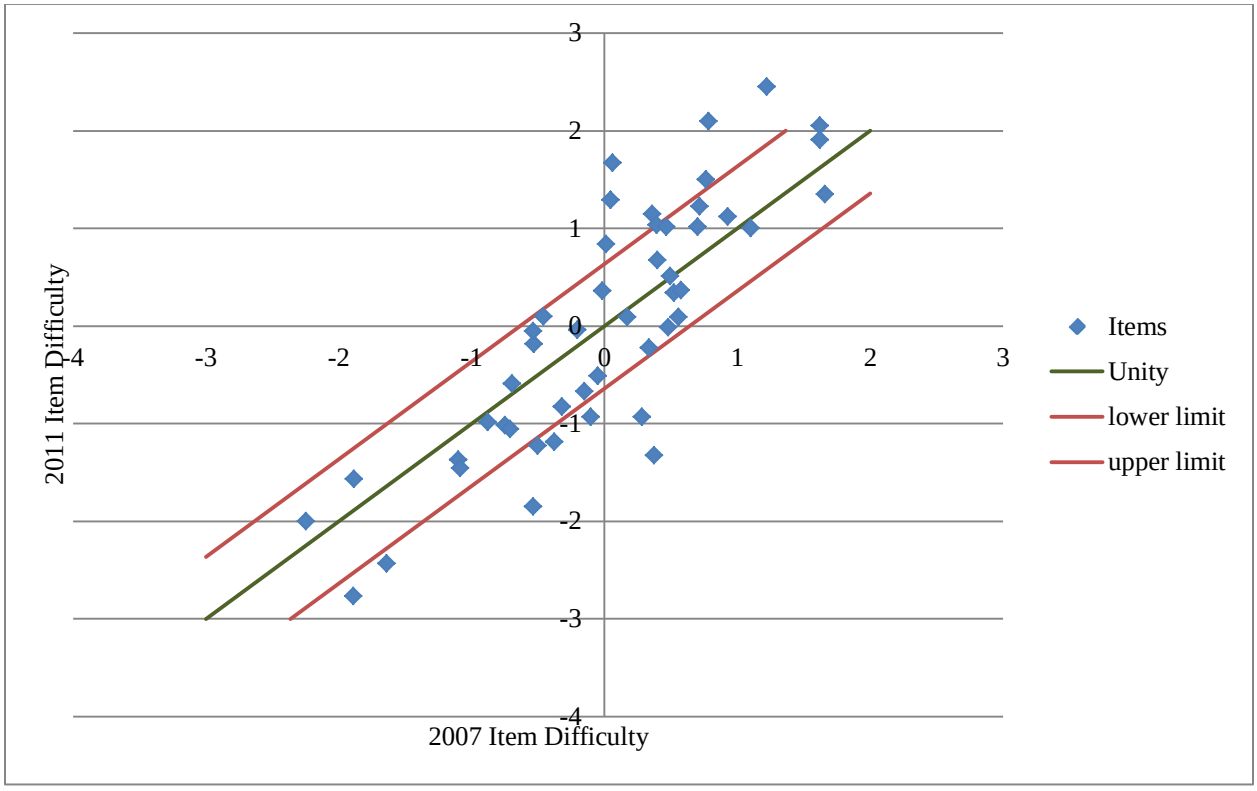


At the item level, 14 and 27 items demonstrated large and statistically significant DIF by language in 2007 and 2011, respectively. Despite the direction of the overall impact, nearly half the items with language DIF favored those who speak official Malagasy and just over half favor a local dialect. Among the items in common for both years, there is substantial variation in which items exhibited DIF. Only 4 items had DIF by language in both years: items 32, 38, 61, and 68. Item characteristic curves for items 32 and 51 by language spoken are shown in figure 2.4. Item 51 (jogging) is an example that favors children who speak official Malagasy (the probability of success is higher for this group as shown with the blue curve in figure 2.4, labeled "Item 51 Model: Malagasy"). Item 32 (goat) is an example of an item that favors a local dialect. There are two words for goat in Madagascar: one used in the central highlands where the official language is spoken, and the other used in the remainder of the country. The latter word was used in the translation and, as a consequence, children who speak official Malagasy found this word

harder (estimated DIF effect is 1.1 and 1.7 logits in 2007 and 2011, respectively). Item 32 also had statistically significant positive infit and this is reflected in the flat red curve (labeled "Item 32 Model: Dialect in figure 3.2") for the children who speak official Malagasy.

As with the mean square fit statistic, I looked for common patterns among the items with DIF. I also consulted with the local Malagasy expert for her opinion. Some interesting patterns emerged. The items that favored children who speak a local dialect include 7 of the 8 French words with no Malagasy equivalent (e.g., panda). The group of items favoring official Malagasy contained some relatively easy vocabulary words that are used commonly in everyday life (e.g. cat, baby, broom, and bottle). However, it is possible that young children are more familiar with synonyms for these words that are used in their local dialect and not the official language. Finally, the group of items without language DIF is interesting in its own right. Two thirds of the items in the last 2 series fell into this group, when children may have been tired by the test. Two of the three words for geometric shapes (i.e., circle, triangle), which were expected to be unfamiliar to all the children also fell into this group. No additional patterns were discerned, although explanations for specific words were found (e.g., the item for "dressing" shows a child putting on socks, but in the coastal areas children walk barefoot or with sandals).

Figure 2.5: Common Item Difficulty by Year - Separate Analyses for subset of 346 children



2.3.3 Multidimensional Model

Anchored Items

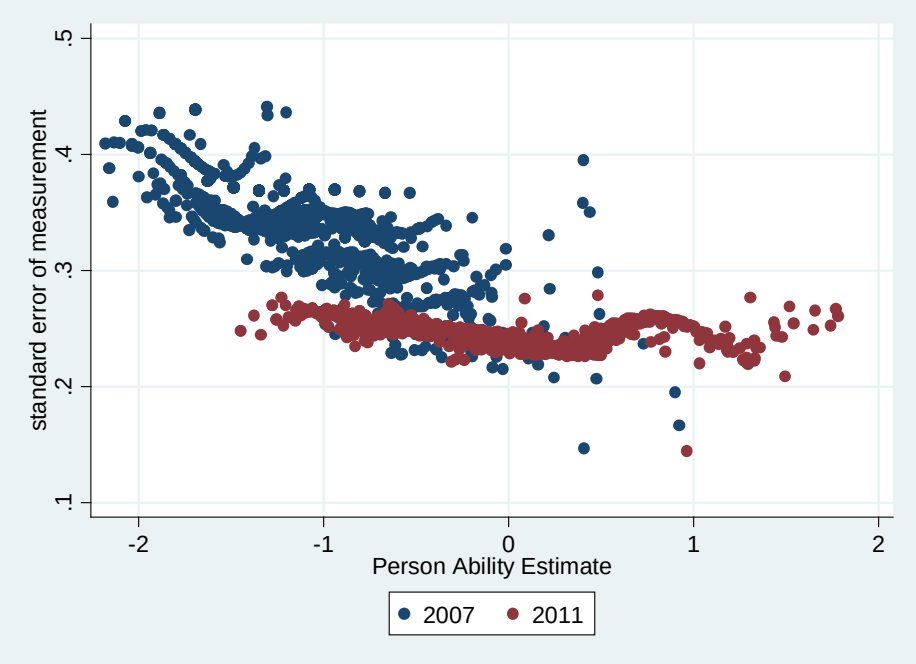
Separate unidimensional calibrations of the 48 common items among the subset of 346 children were found to be acceptable in terms of item fit (not shown) and were much the same as the

results from the full dataset. A plot of the unidimensional item difficulty estimates by year is shown in figure 2.5. Sixteen of the items deviated from a unity line by more than 0.638 logits. These results were replicated with the concurrent analysis of DIF by year of test administration. Out of the 48 items that were administered in both years, 14 suffered from large significant DIF by year of test administration. An additional 4 items had poor item fit statistics when the model was tested on all the children (capturing all 16 from the separate calibrations plus an additional 2 items). Thirty items were kept as anchors and their difficulty constrained for the two-dimensional model.

Reliability & Respondent Measures

The person separation reliability from the two-dimensional model doesn't change substantially from what was achieved with the separate calibrations. The range of ability estimates is also comparable. In 2007, the abilities range from -2.2 to 0.9 logits and in 2011, from -1.4 to 1.8 logits. Not surprisingly, the two-dimensional ability estimates are highly correlated with the raw total scores and uni-dimensional estimates (see table 2.3). However, unlike the other two sets of scores, the average gain in ability from 2007 to 2011 can now be estimated as 0.95 logits (SE 0.017), or about 30% of the full range of abilities in either year. In addition, a dis-attenuated correlation of the two year ability estimates of 0.59 is obtained. This result is comparable to the 0.60 correlation (also corrected for measurement error) obtained from the unidimensional estimates.

Figure 2.6: Standard error of measurement for two-dimensional calibration of '07 & '11



The standard errors of measurement as a function of ability estimates are shown in figure 2.6 for both 2007 and 2011. Estimates of children's abilities in the year where their responses are missing can be obtained from the two-dimensional model, but the standard errors on these estimates are very high (~0.5 to 0.6 logits). Excluding these missing score imputations, the

standard errors of ability estimates in 2007 are shifted downward and reduced on average in the two-dimensional over the unidimensional model (see table 2.2).

Internal Structure

As with the separate unidimensional models, the internal structure of the instrument was checked at the instrument level and at the item level with the use of item fit statistics and Wright maps for the two-dimensional model (see Appendix A1 table 2.5). The distribution of item difficulty estimates is comparable to that of the combined unidimensional calibrations (-4.3 to 2.3). The respondents also have similar distributions of ability estimates, with the 2007 data centered near -1 logits, and the 2011 data centered on the item difficulties at zero. The order of item difficulties shifts somewhat, but still fails to follow the order from the English-language version of the PPVT.

The item difficulty estimates were inspected for all items. As with the separate calibrations, the infit for all of the items are within the acceptable boundaries of 1.33 and 0.75, with many of the same items having statistically significant infit. Specifically, the item with the strongest evidence of negative misfit (t-statistic of -10) once again was item 30 (peeking). The item with the strongest evidence of positive misfit (t-statistic of +8) was again item 86 (tropical). (A plot of the infit mean squares for the items by year is available in Appendix A1 figure 2.13).

DIF by gender and language

The item difficulty estimates obtained from the final multidimensional model were constrained in new unidimensional models and the tests of DIF were repeated separately for 2007 and 2011.

Gender

The estimated difference in overall ability (impact differential) by gender is 0.03 logits (SE 0.017) in 2007 and 0.05 logits (SE 0.011) in 2011. As before, these differences are significant, but small, and girls found the items harder on average than boys. At the item level, 4 of the same 5 items exhibited DIF in 2007 (item 62 DIF is still large, but no longer statistically significant). As with the unidimensional model, the DIF effect for item 72 is large (~1 logit), but it is now statistically significant in the multidimensional model. None of the items had large significant DIF by gender in 2011.

Official Malagasy vs. Local Dialect

Once again, the MD calibrations give results that are very consistent with those obtained from the separate calibrations. The estimated difference in overall ability (impact difference) by language is 0.35 logits (SE 0.019) and 0.65 logits (SE 0.013) in 2007 and 2011, respectively. In 2007, 15 items demonstrated large and statistically significant DIF by language: the same 14 from the unidimensional model, plus one additional item that had moderate and significant DIF previously. In 2011, 27 items demonstrated large and statistically significant DIF; 26 of these were the same as found with the unidimensional model.

2.4 Discussion

2.4.1 Test Validity

Evidence of external validity

The evidence of external validity of the instrument as a whole is consistent with what I would expect from the literature. The raw scores trend upwards with age and correlations with other household and child characteristics are as expected. In both years, there were small but significant differences in mean scores by gender, with boys scoring higher on average (boys and girls have nearly equal enrollment in primary and secondary school in Madagascar).[88] At most, the mean difference between genders was estimated to be 0.11 logits in 2007, or about 3% of the total range of scores in that year. This small differential impact is consistent with a lack of statistically significant differences by gender for the raw totals (all t-test p-values > 0.25).

The mean differential impact by language was significant with approximately a third of a logit difference in ability in 2007, increasing to two thirds of a logit by 2011. Children who speak a local dialect at home (not official Malagasy) scored 10 to 20% lower on average than those who speak official Malagasy. I expected to find some overall difference by language given that children who speak a local dialect are also from poorer households and have less educated mothers. The moderate positive correlation of the scores with mother's education and household wealth support this expectation. In addition, the increase in the differential impact from the younger to the older cohort is in keeping with reports of a socio-economic gradient for cognitive outcomes (including the PPVT) that widen as children age.[48, 89]

Evidence of internal validity

The evidence of internal validity of the instrument is less favorable. Of the 96 items administered in either year, 53 (55%) demonstrated some evidence of poor fit, either from statistically significant mean square error (infit) or DIF (by gender or dialect) or both. The lack of fit explains why such a large number of items (i.e., 72 items) were necessary to obtain a person separation reliability of over 0.8 in 2011, and to keep the standard error of person ability estimates low (under 0.3 logits). The problem of items with positive infit from test fatigue can only be resolved by reducing the length of the test, which in turn can only be accomplished by re-ordering the items. The problem of items with negative infit related to high discrimination can only be resolved by removing, replacing, or editing the items (see discussion on recommendations).

Only six items exhibited gender DIF in 2007 (none in 2011), and four of these favored boys over girls. Although I cannot conclude whether the differential impact seen is due to a true difference in abilities or item-level bias, the differences are small enough that I did not explore them further. On the other hand, 37 (38%) of the 96 items exhibited DIF by language. The items are split in terms of whom they favor, so again I cannot rule out the possibility that the mean impact may be partially due to item level bias. The item DIF effect sizes are also quite large, with 9 items in each year exceeding one logit (~30% of the full range of scores). The item characteristic curves for item 32 (goat) are particularly telling (figure 2.4). The observed responses from children who knew the word (from the lowlands) follow the modeled probabilities very well. On the other hand, the observed responses from the children who didn't know the word (from the highlands) have a flat curve indicating positive infit. And in fact, the most commonly selected distractor for item 32 was the top right hand image #2.

At first the direction of DIF among the French words seemed counter-intuitive. I presumed that knowledge of French would be associated with wealth and living in the central highlands near the capital (where they speak the official Malagasy). But there are some possible explanations that suggests otherwise. The local expert with whom I consulted suggested that many of the coastal dialects incorporated French words or foreign-sounding words into their vocabulary when Madagascar was colonized. In addition, I found a reference that states: "instruction in French is preferred by the coastal peoples, as it avoids connotations of Merina cultural dominance."[75] In my sample, nearly all the children on the coast were categorized as speaking a local dialect. Future analysis will include a more detailed evaluation of the DIF by geographic region or ethnicity.

2.4.2 Use of Multidimensional Model

This chapter describes the use of a multidimensional item response model to estimate children's vocabulary knowledge for repeated measures at two time points, 4 years apart. The model is based on a latent growth item response model (LG-IRM) where the "gain" over the two time points is estimated. The model allows for direct comparisons of abilities by creating a common scale from a set of 30 items with constrained item difficulties. Unlike the standard approach of separate estimations to obtain ability estimates for each year, the multidimensional approach provides better estimation accuracy of abilities in both groups by using collateral information one from the other. I was able to reduce the standard errors for abilities in 2007 by borrowing information from 2011, and obtain ability estimates for children with missing scores in one of the years. Finally, having the two time points on the same scale, allows me to estimate the change in ability over time (0.95 logits or 30% of the full range of abilities in a given year).

Both the unidimensional and multidimensional models identified almost exactly the same set of items with poor infit and DIF. These results suggests that information on item fit is stable over time and not enhanced by combining the data from the two years into a concurrent analysis. Although the multidimensional model provides important advantages to the estimation of person ability, it does not appear to be necessary for identifying problem items in the test. The simpler unidimensional model may be sufficient during pilot testing to improve the performance of the instrument before full test administration.

2.4.3 Recommendation

The following recommendations are based on the main take home points that I have gained from the research described in this chapter. The recommendations are specific to a situation where an existing instrument has been carefully translated and pre-tested, in collaboration with local experts, for use in a setting for which it was not originally designed. I split my suggestions into pre-test (i.e., pilot phase) and post-test (i.e., analysis phase) administration. Throughout this section, I recommend the use of IRT methods, but in some cases I am able to suggest alternatives using classical test approaches. However, these alternatives lack the detailed statistics available from IRT for making decisions about keeping, dropping, or re-ordering items. They also lack information about the standard error of person ability obtained with IRT that can inform whether any individuals should be dropped from the analysis or if enough items are being administered.

Pre-test administration

If significant modifications to the instrument are allowed by the publisher, then a number of steps can be taken after piloting to minimize bias in the final test scores (note that the publisher for the PPVT currently does not allow these types of changes without permission). First, problem items need to be identified by pre-testing many more items than are thought to be necessary, without the use of stopping rules. Items with strong evidence of poor fit can either bias the overall score (from DIF) or contribute little to the estimation of person ability. For example, if many items are too highly discriminating, then they are redundant and make the test longer than necessary. Items with positive misfit suggest a non-random pattern of responses worth investigating. In the absence of IRT methods, some item misfit can be roughly identified by ordering the items and persons by total correct number of responses in a simple Excel spreadsheet. If multiple children with low total scores are getting certain hard items correct, then this may be evidence of positive misfit of the item (or interviewer bias). Local experts may be able to accurately identify words that are likely to cause DIF (i.e., the two words for "goat" in Madagascar).

Once problem items are identified, they can be removed, replaced or modified. Replacement may only be feasible if not all the items were used in pre-test or the test has multiple forms (i.e., the PPVT is available as form A or form B). The simplest way to modify a problem item is to change the image to be more culturally relevant (e.g., this was done in Madagascar for the word "money") or to try an alternate translation to the stimulus word. Another possible modification is to switch the stimulus word to one of the distractor images that is not linguistically or otherwise thought to be biased (e.g., this was done for an adaptation into Greek).[90]

Next the new set of items needs to be re-tested and the items re-ordered from easiest to hardest. If necessary, the spreadsheet method can be used to re-order the items. This is likely to be an iterative process. At the final stage, stopping rules can be implemented to avoid test fatigue. I hesitate to recommend using different starting points for different ages for two reasons. First, I have found that the rules for starting mid-test and finding the basal set are difficult to train and implement properly in large scale studies. Second, the exact age of the respondent may not be known, and selecting the wrong starting point introduces problems for scoring. However, in studies with a wide range of ages, different starting points may be necessary.

If significant modifications to the instrument are not allowed by the publisher, then only a couple of steps can be taken to limit exposure to possible problems in the analysis phase from the test administration. Again, I would start with pre-testing many more items than are thought to be necessary, without the use of stopping rules. Identify the minimum set of consecutive items that will work with the age group being tested for the best person separation reliability. Administer this same set of items to the full sample without start or stopping rules. If a wide range of ages is being tested, then two sets of consecutive items can be identified, with some overlapping items between the groups. For example, in 2011, 6 series of 12 items (72 items) was used for children 7-10 years of age, and this series was extended by 2 additional series for adults.

Analysis Phase

There are at least four options for handling poor fit or DIF at the item-level in the analysis phase. First, the items can be ignored. This is probably the most common option researchers choose

either because a) a detailed item-level analysis is beyond the scope or familiarity of the researchers (IRT methods are still fairly new), or b) it is unclear what option would be better. If the DIF effect happens to balance out in the summary score, the consequences of choosing to do nothing may be nil. On the other hand, if the DIF effect favors one group over another, then the ability estimates will be biased. The consequences are then a biased program effect estimate and possibly drawing the wrong conclusion from the study.

A second option is to drop the items that exhibit poor fit or DIF. There is precedence for this approach. For example, researchers from the Young Lives study reported finding items with gender DIF after data was collected for the study. They excluded these from their analysis.[76, 77] Although this second option may seem reasonable because item-level bias is eliminated, there are trade-offs to consider before doing so. In the Madagascar dataset, the person separation reliability in 2007 is only moderate (~0.64). Dropping items might reduce the reliability to low levels. In addition, some of the DIF that I found is likely due to chance alone, and items would be dropped unnecessarily. Finally, although I would gain one form of validity (at the item level); I might lose another form of validity (at the instrument level). Specifically, the underlying construct being measured may no longer be the same. This is also true if items are dropped or replaced pre-test administration. The instrument will have shifted to measuring something new - something that reflects no DIF by group membership.

A third option is to perform separate analyses by subgroup. The investigators of the Young Lives used this option when one or more items were flagged with language DIF. The loss of power is one obvious drawback. But more importantly, this approach changes the research question.

The final option is to incorporate an interaction term for the item and group membership into the item response model when estimating person ability. This is probably the best option, although the interpretation of person ability becomes complicated as the number of subgroups increase. In this chapter, I only investigate differences by gender and a dichotomous indicator for dialect. However, there are many dialects in Madagascar (with the same base syntax and about a 70% overlap in lexicon),[75] and this information was not captured in the 2011 survey. The dialect indicator is a catch-all that is just as likely to reflect geographic and ethnicity differences as it does language (the official Malagasy is closely related to the Merina dialect spoken in the central highlands among those of Malayo-Polynesian descent as opposed to those of African descent living on the coast).

2.4.4 Conclusions

Why use IRT

The use of IRT is not common in the health field and I have found limited evidence of its use for evaluating the validity of language instruments that are used out of cultural context of their original design and in a different language. However, I demonstrate the necessity of validating a translated, multi-item instrument when it is used in a context for which it was not developed. Importantly, I show how item response methods can uncover problem areas and patterns of responses that would not be apparent with the classical test theory approach, especially at the item level. For example, I was able to identify sources of item-level bias by statistically testing for the presence of DIF by language subgroup. Evidence of test fatigue observed in the field was confirmed by item infit statistics. Infit statistics also provided evidence that was contrary to

my expectations about the direction of bias from using French words. In addition, by placing both person ability and item difficulty on the same mathematical scale, I was able to visualize (with the Wright maps) how well the item difficulties cover the full range in person abilities (e.g., whether the items were too hard on average).

A very important lesson was learned from the use of IRT analyses: beware the use stopping rules when administering a test whose order of item difficulties may be lost in translation! The loss of order, combined with the use of stopping rules, resulted in censoring and uncertain ability estimates for a large proportion of the children in 2007. They were administered many fewer items and were prevented from taking easier items that appeared in later series. Although incredibly useful, the 2011 work-around of not using stopping rules was less than ideal. The longer test appears to have resulted in fatigue. In future studies, the estimated difficulties from IRT can be used to re-order items, allowing for the re-introduction of stopping rules (with the publisher's permission).

The fact that IRT models do not rely on an automatic scoring rubric of non-administered items is a clear benefit over the classical approach of estimating abilities, especially in a situation with censoring as just described. The IRT probability model relies on the responses to items by subjects to inform the ability of those who were not administered the items, even in a single year's administration. By borrowing information from 2011 in a multidimensional model, I obtain an even better estimate of ability (smaller standard error) for children who were censored in 2007. This sharing of information across respondents and across surveys is not possible with the classical approach.

Future research

The next step in my research of the PPVT results in Madagascar is to obtain the least biased estimate of children's abilities, addressing the evidence from problematic items. Items with negative infit will be left in the estimation process – they do no harm. A cut-point or specific criteria will be chosen for dropping the items with the strongest evidence of positive infit and the largest DIF effect size for language. The criteria will be set so as to minimize the loss of items and the impact on reliability. Any remaining DIF will be handled by incorporating interaction terms into the model. Once I've obtained ability estimates that I am confident with, these will be used to evaluate whether the Madagascar nutrition program had an effect on children's vocabulary knowledge. Note that this requires the use of plausible values (or imputed values), as opposed to the IRT estimates, if the evaluation is run outside of Conquest. I will be able to further evaluate the advantages of using the IRT methods discussed herein by comparing estimates of a program effect with and without its use.

Chapter 3: Identification of the Target Parameter – the Case of Pre/Post Data

3.1 Introduction

In the previous chapter, I presented the measurement challenges faced by researchers who want to evaluate program effects on measures of early cognition, specifically vocabulary knowledge. I discussed several sources of bias and offered suggestions of how to obtain a minimally biased estimate of language ability from the measure. Presuming that I have an outcome assessed without bias, the next step is to evaluate whether the intervention had an effect on this outcome. In this chapter, I illustrate the application of the first part of an analytic framework, or road map, for evaluating the population average treatment effect (ATE) of a program.[58] These steps include the process of defining the target causal parameter and stating the assumptions under which it is identified as a parameter of the distribution of the observed data (discussed in detail in the methods section).

To demonstrate the use of the road map in the evaluation of an intervention, I again make use of the program in Madagascar as a backdrop for exploration. As with the language measure, the Madagascar program presents some interesting challenges, all of which are common in impact evaluations of ECD interventions. First, the program was implemented at the community rather than the individual level. Community programs differ from individual treatment in that they are typically made available to all (or most) residents of a community. When the point of treatment is shifted to the community, the research question shifts to what would happen to the community under a given treatment assignment. In this context, I need to reframe the question, building up by analogy, from the individual to the group. I use the potential outcomes framework (also known as the counterfactual framework) popularized by the work of Rubin to help in this regard.[91] In a counterfactual framework, I consider the "ideal experiment" when posing my research question: for example, what would have happened to a given community (instead of an individual) had it received treatment (the counterfactual) when in fact it had not?

Second, the program rollout was non-random: communities with the greatest perceived need were targeted first. This non-random assignment of treatment to the poorest communities makes inferences about the programs' effect susceptible to confounding, if the comparison group is not exchangeable with the treated group on key determinants of the outcome. Therefore, it is imperative to define a causal model for the system that is hypothesized to have generated the data and to examine clearly the relations and dependencies of the factors in the model (measured and unmeasured). I avoid imposing restrictions on the functional form of these relationships by using semi-parametric structural equations and directed acyclic graphs (DAGs).[92, 93]

Third, cross-sectional surveys were administered in the same communities in Madagascar pre- and post-intervention, providing multiple options for identification of a causal effect. In this chapter, I contrast the definition of the outcome as either: a) the post treatment value or b) the change from pre to post treatment. I consider the long-standing controversy over the advantages and disadvantages of each.[60, 94] Using these two outcomes, I identify three statistical parameters (commonly used for interventions with pre-post data) that under different assumptions are equivalent to my causal target parameter of interest, the ATE. These parameters include a post treatment estimand (adjusting for the lagged outcome) and two difference-in-differences models:

a change score estimand and a pooled outcome estimand (popular in the social sciences and econometrics). [60-63] I highlight the assumptions imposed by each of these models. Using data simulations, I show how the causal effect estimate is biased when these assumptions fail to hold.

3.2 Methods

3.2.1 Setting

In Madagascar, treatment assignment was made in such a way that communities with the greatest perceived need were to receive the program first based on outcome(s) measured pre-implementation and other logistical factors. The outcome of interest is a measure of nutritional status in children that can be measured with minimal error given adequate training (e.g., weight-for-age). Data are obtained with repeated cross-sectional surveys, pre and post implementation of the program, from both program participating and non-participating communities. Each sample includes different children, but the same communities. The type of data collected is described in general terms in table 3.1.

3.2.2 Notation

I use the notation shown in table 3.1 throughout this chapter and the next (based on the book on *Targeted Learning* by van der Laan and Rose).[58] The notation has been modified to indicate how I aggregate individual level measures up to the community level (e.g., mean maternal education), and that these are different from community variables that are measured once for the entire group (e.g., geographic location).

Table 3.1: Notation

V	Vector of time invariant community level covariates (e.g., urban location)
$W^c(t)$	Vector of community level covariates that summarize individual level factors, $W_i(t)$, (i=1,…,N), for each of the N individuals sampled in the community at time t=0,1 (e.g., proportion of mothers sampled in the community who are uneducated)
A	Treatment, assigned at the community level
$Y^c(t) = \frac{1}{N}\sum_{i=1}^{N} Y_i(t)$	Community mean of individual level outcomes $Y_i(t)$ (i=1,…,N) for each of the N individuals sampled in the community at time t=0,1 (e.g., weight-for-age of children under 5 years)
$O_j = (V_j, W^c_j(t), A_j, Y^c_j(t))$	Observed data structure, O_j, for a given community j. The observed data are J independently and identically distributed copies of O.
$U_V,\ldots U_{Y(t)}$	Random variation for each variable
P_0	True data-generating distribution; $O_j \sim P_0$
Y^c_a, Y^θ_a	Counterfactual outcomes; I focus on 2 outcomes: post treatment outcome, $Y^c(t=1)$, and the change in outcome pre and post treatment, $Y^\theta = Y^c(t=1) - Y^c(t=0)$. For each, I define their counterfactual value under treatment level A=a (Y^c_a and Y^θ_a, respectively).
$\Psi(P_0)$	True value of the target statistical parameter (or estimand), consisting of parameter mapping Ψ applied to the true data generating distribution P_0. I present 3 estimands labeled Ψ^I, Ψ^{II}, and Ψ^{III}.

3.2.3 Causal Inference Road Map

The road map I follow links the research question to inferences from the results, making the underlying assumptions explicit along the path between the two (see van der Laan and Rose, chapters 1 and 2 for more detail).[58] First, I define precisely the research question. This may seem obvious, but is often not made clear at the outset. Second, I turn the research question and relevant background knowledge into a structural causal model (SCM), which encodes information about the relationships between the variables. Importantly, I assume that the SCM accurately represents the data generating processes that gave rise to my observed data. This is the key link from counterfactual to observed data. I use a semi-parametric variant of a structural equation model to avoid making assumptions about the underlying functional form of the data distribution.[93]

Given the SCM, I specify the causal parameter of interest in the third step. The causal parameter is the parameter I would obtain under an ideal experiment and is defined using counterfactual notation. A clear specification of the causal parameter requires an understanding of a) the outcome (i.e., a post treatment value or a pre-post change score); b) the variable or variables on which I want to intervene (i.e., program availability or program participation) and the unit (or level) on which I am intervening (i.e., the individual or the community); and c) which counterfactual outcome distributions (or parameters of these distributions) I want to compare. In this chapter, I use phrasing such as "intervening to set the treatment" or "setting A=a" to refer to the hypothetical treatment condition that I want to apply to the system when making causal contrasts. For example, I might be interested in estimating the difference in the expectation (or mean) of counterfactual outcomes "intervening to set the treatment" to 1 (to receive treatment) versus "intervening to set the treatment" to 0 (to not receive treatment) for all communities. This contrast is known as the average treatment effect, or the ATE. Alternatively, I might want to evaluate the average treatment effect among the treated, or the ATT. The ATT contrasts the expectation of counterfactual outcomes under treatment and no treatment, but only among the treated communities. Importantly for both causal parameters, the contrast is made between the means of the counterfactual outcomes under each treatment regime. This is a simpler causal comparison than between the two potential outcomes for any given community, where one outcome is always unobserved.

In the fourth step, I assess identifiability, or whether the observed data, in combination with my assumptions about the data generating system, are sufficient to express the target casual parameter of interest as a parameter of the distribution of the observed data alone. This second parameter is the statistical target parameter (also referred to as the estimand; I use the terms interchangeably in the text). In contrast to the causal parameter, the estimand is the parameter that I am actually able to estimate given the observed data. In this step, I make use of the SCM to carefully evaluate the assumptions for each of three estimands. In the first estimand, the outcome is defined as the outcome post treatment ($Y^c(t=1)$); in the second, the outcome is defined as the change in outcome pre- vs. post-intervention (Y^θ); and in the third, the outcome is pooled over time ($Y^c(t)$). In addition to encoding the causal model as a series of equations, I depict the same information as a directed acyclic graph (DAG). The advantages of using DAGs are discussed in more detail (and become apparent) in the identifiability section.

In the last steps, I can commit to an estimand and statistical model and proceed with the estimation. However, in this chapter, I use simulations to illustrate the different assumptions required for the three statistical parameters to be equivalent to my target parameter of interest, and the consequences when the assumptions do not hold. In chapter 4, I will present estimation and inference results for the observed data from Madagascar. I present detailed steps 1 through 4 of the road map next.

Steps 1-3: The Research Question, Target Causal Parameter & SCM

The causal question that I want to answer is: Does the intervention increase the average nutritional status of children living in the community? I am interested in estimating a population average effect at the community level, for all communities in the target population.

The structural causal model (SCM) is characterized by a set of endogenous variables at two time points (see notation table 3.1). Community variables that are not aggregates of individual factors are denoted by V, and are assumed to be time-invariant for the period of the study. Individual level factors aggregated up to community-level factors are denoted by a vector, $W^c(t)$, at time t. The community-level mean outcome for children at time t is denoted as $Y^c(t)$. The community level exposure, A is assigned to zero or one at t=1 as a function of V, $W^c(t=0)$ and $Y^c(t=0)$. In addition, there are unmeasured exogenous variables, U, that may cause random variation in each of the observed variables. Restrictions on the joint distribution of these unmeasured errors will be required for identifiability, which I discuss below.

I pose the following structural causal model (SCM) to explain the relationships between the variables:

$V = f_V(U_V)$
$W^c(t=0) = f_{W(t=0)}(V, U_{W(t=0)})$
$Y^c(t=0) = f_{Y(t=0)}(V, W^c(t=0), U_{Y(t=0)})$
$A = f_A(V, W^c(t=0), Y^c(t=0), U_A)$
$W^c(t=1) = f_{W(t=1)}(V, W^c(t=0), Y^c(t=0), U_{W(t=1)})$
$Y^c(t=1) = f_{Y(t=1)}(V, W^c(t=0), Y^c(t=0), A, W(t=1), U_{Y(t=1)})$,

where no assumptions are made about the shape or form of the functions. I start with a model with a minimal set of exclusion restriction assumptions about the data-generating system in order to avoid imposing restrictions that may, or may not be, supported by the data. I make a single exclusion restriction in this model: that the covariates $W^c(t=1)$ occurring post intervention are not affected by the intervention. I impose this deliberate exclusion restriction for three reasons. First, it is a reasonable assumption in the context of the Madagascar study. Second, it is required for the estimand with the outcome pooled over time (see identifiability section for estimand III). I apply this restriction to the other two estimands in this chapter to facilitate my comparison among them (although it is not required). Finally, it allows me to condition on $W^c(t=1)$ in the models to better predict $Y^c(t=1)$ (also not required for the first two estimands).

In the ideal experiment, I would want to know what would happen to the population mean outcome, if every community had the program, versus none of the communities had the program. I translate this into my target causal parameter as the average treatment effect (ATE) given by:

$E(Y^c_1(t=1) - Y^c_0(t=1))$, where $Y^c_a(t)$ denotes the counterfactual community level outcome under an intervention on the SCM setting A=a. In the next step, I describe three identifiability results (and corresponding estimands) where I link this causal parameter to my observed data distribution.

Step 4: Assess Identifiability

Causal effect estimation relies on assumptions, some of which cannot be tested. These assumptions must be made explicit when using observational data for causal inference. Specifically, the identifiability of my causal target parameter requires some form of the following two assumptions to hold: the randomization assumption (RA) and the experimental treatment assignment (ETA) assumption.

The RA (also known as the assumption of no unmeasured confounders, or of exchangeability), states that treatment, A, is independent of counterfactual outcome, Y_a, given some subset of the data. The RA is a causal assumption, and as such is not testable. However, I can draw a graphical representation of my SCM (i.e., a DAG) to check the independence assumptions given my knowledge of the underlying data generating system.[92] By using a graphical procedure, I am able to solve the identification problem without resorting to an algebraic analysis of whether a statistical model parameter has a unique solution in terms of the parameters of the distribution of the observed variables. [93] Detailed guidelines for reading causal diagrams are available in *An Introduction to Causal Inference* by Judea Pearl,[93] or in *Causal Diagrams for Epidemiological Research* by Sander Greenland et. al.[95] Very breifly, the graph is drawn based on the relationships defined in the SCM, where the parents of a variable (variables on the right hand side of the equation) are connected to the child variable (variable on the left hand side of the equation) with an arrow directed towards it. A path is any sequence of lines connecting two variables. The arrow between two variables can only go in one direction, such that the paths are acyclic (i.e., the graph cannot have $A \rightarrow B \rightarrow C \rightarrow A$). Paths can either be open or blocked, depending on the direction of the arrows and whether or not a variable is conditioned on. Conditioning on a variable is represented by placing a box around it. Open paths can give rise to dependency between variables, and the absence of any open paths implies mariginal indepenence.

The specific randomization assumption (RA) and necessary additional assumptions for three different estimands are discussed in detail below. To minimize confusion from too many arrows, I represent DAGs for each estimand using a simplified data structure that omits the observed, time-invariant, village factors, V. I can justify this simplification because V are exogenous to the data generating system (no arrows go into V) and if I condition on V, I do not have worry about unblocked paths from unmeasured variables through V. In most cases, I also omit the exogenous variables, U, such that $O_j = (W^c(t), A, Y^c(t))$. The omission of the U's implies that these exogenous variables are independent (discussed further with figure 3.1). Paths depicted in red in the figures represent unblocked paths between the treatment and outcome variables.

The ETA assumption (also known as the positivity assumption) states that for the target statistical parameter to be identified there must be sufficient variation in treatment (i.e., some positive probability of both being treated and not being treated) within strata of confounders. The form of the ETA assumption depends on knowledge of the data-generating system encoded

in the SCM and on the target parameter. For the average treatment effect, the strong positivity assumption states that each possible treatment level occurs with some positive probability within each stratum of the confounders.[96] But this can be weakened under additional parametric assumptions. For example, urban versus rural location is a confounder in my study in Madagascar. The strong version of the ETA assumption requires that I have both treated and untreated, urban and rural communities in my observed data. If, in fact, there were no observed treated urban communities, then I could weaken the ETA assumption by assuming (if plausible) that the treatment effect is the same among urban and rural communities. However, imposing this type of parametric assumption is risky as it requires extrapolating from an area supported by the observed data (the treatment effect among rural communities) to an area that is not (the treatment effect among urban communities).[96] The ETA assumption will be discussed in more detail in chapter 4 in the context of the actual data from Madagascar. For the purposes of this chapter, I accept that the ETA assumption is not violated in my study.

There are two additional assumptions that are typically invoked when investigators start from the Rubin framework of potential outcomes for causal inference: the consistency assumption and the stable unit treatment value assumption (SUTVA).[64] Both assumptions are subsumed in my SCM and the implied knowledge it encodes about the underlying data generating distribution. The consistency assumption states that an individual's (or community's) potential outcome under the treatment actually received is precisely the observed outcome.[64] This assumption is used to convert probabilities written in terms of counterfactuals into ordinary probabilities in terms of observed values. However, my SCM already implies the counterfactual and provides the necessary link to the observed data. In addition, the absence of hierarchical relationships between communities in my SCM implies that one community's (or individual's) outcome is unaffected by another's treatment assignment (i.e., SUTVA holds).

Estimand I: Outcome $Y^c(t=1)$

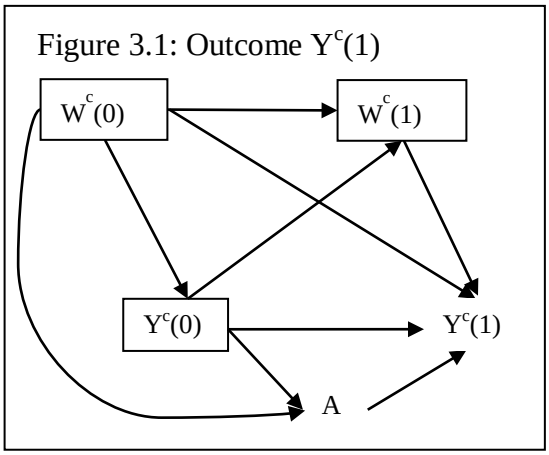
For the first estimand, I define the outcome as the community specific mean post-treatment outcome, $Y^c(t=1)$. Identifiability is based on conditioning on all baseline covariates, including the pre-treatment (or lagged) outcome (as well as the post treatment covariates $W^c(t=1)$ assumed not to be affected by A, as discussed above). This is a common approach in the epidemiology literature. The RA for this estimand is:

$$Y^c_a(t=1) \perp A \mid V, W^c(t=0), Y^c(t=0), W^c(t=1) \qquad (1)$$

For the randomization assumption (1) to hold, it is sufficient that the exogenous variables for the exposure, U_A, be independent of the exogenous variables for the outcome, $U_{Y(t=1)}$, given V, $W^c(t=0)$, $Y^c(t=0)$, $W^c(t=1)$. This additional independence assumption is reasonable, if I have no unmeasured common causes of A and $Y^c(t=1)$ (i.e., no confounders).

The DAG in figure 3.1 encodes the information from the series of equations from the SCM in the previous section. The graphical model is particularly useful in that I can visually check that $Y^c(t=1)$ is independent of A given $W^c(t=0)$, $Y^c(t=0)$, and $W^c(t=1)$. Specifically, I check that my conditioning variables block any unblocked path from A to $Y^c(t=1)$ (i.e., paths with arrows pointing into A), while not opening any new paths. This is referred to as the backdoor criterion.[95] In figure 3.1, the variables $W^c(t=0)$, $Y^c(t=0)$, and V (not shown) are conditioned on,

block the paths from A to $Y^c(t=1)$, and satisfy the backdoor criterion. Writing the graph in this way implies the independence assumptions among the exogenous variables, U, described previously. The RA(1) holds under this model.



I now have the following identifiability result:

$$E[Y^c_a(t = 1) \mid V, W^c(t=0), Y^c(t=0), W^c(t=1)] = E[Y^c_a(t = 1) \mid A=a, V, W^c(t=0), Y^c(t=0), W^c(t=1)] = E[Y^c(t = 1) \mid A=a, V, W^c(t=0), Y^c(t=0), W^c(t=1)]$$

where the first equality holds under the RA(1), and the second holds under my definition of the counterfactual outcomes. Note that for these conditional expectations of the outcome to be well-defined in my SCM, I need some communities with and without the treatment for each level of the conditioning variables V and $W^c(t)$ (i.e., I need for the positivity assumption to hold).

A first estimand (or statistical parameter) for the average treatment effect, Ψ^I, follows:

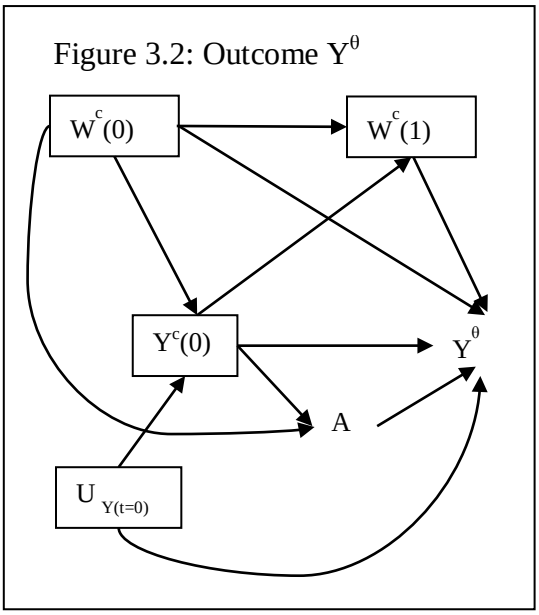
$$\Psi^I(P_0) = E_{V,W(t=0),Y(t=0),W(t=1)}[E[Y^c(t = 1) \mid A = 1, V, W^c(t = 0), Y^c(t = 0), W^c(t = 1)] - E[Y^c(t = 1) \mid A = 0, V, W^c(t = 0), Y^c(t = 0), W^c(t = 1)]] \quad (2)$$

I refer to this estimand as the post treatment estimand.

Estimand II: Outcome Y^θ

Next, I consider the outcome as the change in the community specific means, Y^θ, before and after treatment. I define Y^θ as:

$$Y^\theta = Y^c(t=1) - Y^c(t=0) \quad (3)$$



By definition of the structural equations for $Y^c(t=1)$ and $Y^c(t=0)$, I have the following structural equation for Y^θ:

$$Y^\theta = f_{Y(t=1)}(V, W^c(t=0), Y^c(t=0), A, W^c(t=1), U_{Y(t=1)}) - f_{Y(t=0)}(V, W^c(t=0), U_{Y(t=0)})$$

The DAG in figure 3.2 reflects this same information. Note that $U_{Y(t=0)}$ now affects both $Y^c(t=0)$ and Y^θ, so I have included it in the graph. Under this model, I have a new RA for outcome, Y^θ:

$$Y^\theta_a \perp A \mid V, W^c(t=0), Y^c(t=0), W^c(t=1) \quad (4)$$

and I can identify a statistical target parameter based on Y^{θ} that is equivalent to Ψ^{I}.[97] Specifically, if I define the counterfactual mean of Y^{θ}_{a} under an intervention on the SCM setting A=a as:

$$E[Y^{\theta}_{a}] = E[Y^{c}_{a}(t=1) - Y^{c}_{a}(t=0))] = E[Y^{c}_{a}(t=1)] - E[Y^{c}_{a}(t=0)] \quad (5)$$

then I can rewrite my target causal parameter in terms of Y^{θ}_{a}, and show that it is identical to the ATE as previously defined as $E[Y^{c}_{1}(t=1) - Y^{c}_{0}(t=1)]$. First, the parameter is expressed as a difference in the differences of means:

$$E[Y^{\theta}_{1} - Y^{\theta}_{0}] = (E[Y^{c}_{1}(t=1)] - E[Y^{c}_{1}(t=0)]) - (E[Y^{c}_{0}(t=1)] - E[Y^{c}_{0}(t=0)]) \quad (6)$$

However, since intervening to set the treatment cannot affect the pre-treatment outcome ($Y^{c}_{a}(t=0) = Y^{c}(t=0)$), the above can be rewritten such that the mean of $Y^{c}(t=0)$ cancels out to give the ATE:

$$(E[Y^{c}_{1}(t=1)] - E[Y^{c}(t=0)]) - (E[Y^{c}_{0}(t=1)] - E[Y^{c}(t=0)]) = E[Y^{c}_{1}(t=1) - Y^{c}_{0}(t=1)] \quad (7)$$

Under the RA (4), I can identify my statistical target parameter

$$E(Y^{\theta}_{a} \mid E, W^{c}(t=0), Y^{c}(t=0), W^{c}(t=1)) =$$
$$E(Y^{\theta}_{a} \mid A=a, E, W^{c}(t=0), Y^{c}(t=0), W^{c}(t=1)) =$$
$$E(Y^{\theta} \mid A=a, E, W^{c}(t=0), Y^{c}(t=0), W^{c}(t=1))$$

and have an alternative, but equivalent, formulation of estimand Ψ^{I}:

$$\Psi^{I^{*}}(P_0) = E_{E,W(t=0),Y(t=0),W(t=1)}((E(Y^{\theta} \mid A=1, E, W^{c}(t=0), Y^{c}(t=0), W^{c}(t=1)) - (E(Y^{\theta} \mid A=0, E, W^{c}(t=0), Y^{c}(t=0), W^{c}(t=1))) \quad (8)$$

So what is the advantage of using Y^{θ} over $Y^{c}(t=1)$ for estimating the ATE? The main justification in the causal inference literature is that a difference method allows for both the treatment, A, and outcome, $Y^{c}(t)$, to depend on unobserved community fixed effects that are time invariant.[60, 98] To explore this advantage, I add an unmeasured confounder, $C = f_{C}(U_{C})$, to my SCM and DAG, such that C is a common cause for A, $Y^{c}(t=0)$, and $Y^{c}(t=1)$ (see figure 3.3). The allowed functional forms of $f_{Y(t=0)}$ and $f_{Y(t=1)}$ in the SCM are restricted such that C has a linear additive effect on $Y^{c}(t)$, specifically that:

$$Y^{c}(t=0) = f_{Y(t=0)}(V, W^{c}(t=0), U_{Y(t=0)}) + C$$
$$Y^{c}(t=1) = f_{Y(t=1)}(V, W^{c}(t=0), Y^{c}(t=0), A, W^{c}(t=1), U_{Y(t=1)}) + C$$

The introduction of an unmeasured confounder, C, opens up a backdoor path from A to $Y^{c}(t=1)$ (see path $A \leftarrow C \rightarrow Y^{c}(t=1)$ labeled (i) and colored red in figure 3.3). The RA(1) for estimand I no longer holds. At first, it appears that RA(4) might hold for Y^{θ}. If I assume C has a constant additive effect on both $Y^{c}(t=0)$ and $Y^{c}(t=1)$, then Y^{θ} is not a function of C when taking the difference of Y^{c} at the two time points. The structural equation for Y^{θ} remains as:

$Y^{\theta} = Y^{c}(t=1)-Y^{c}(t=0) = f_{Y(t=1)}(V,W^{c}(t=0), Y^{c}(t=0), A, W^{c}(t=1), U_{Y(t=1)}) - f_{Y(t=0)}(V,W^{c}(t=0), U_{Y(t=0)})$

Figure 3.3: Outcome Y(1), unmeasured C

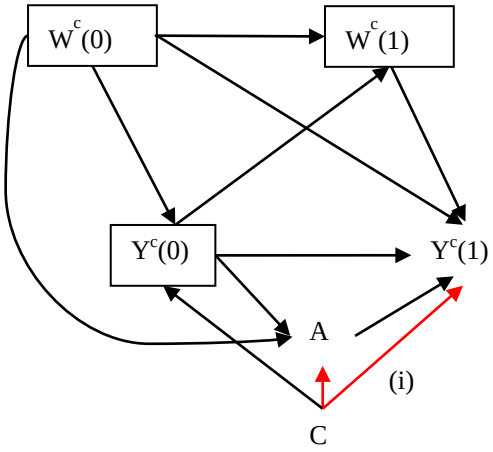
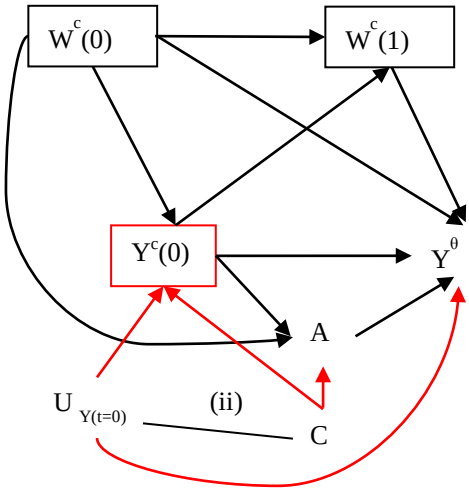


Figure 3.4: Outcome Y^{θ}, collider at Y(0)



The DAG for Y^{θ} in figure 3.4 reflects this same information in that there is no arrow from C into Y^{θ} (only variables on the right hand side of the equation have arrows into Y^{θ}). Thus using Y^{θ} instead of $Y^{c}(t=1)$ as outcome has the potential (under this specific parametric assumption) to close one backdoor pathway from A to Y^{θ} via unmeasured confounder C.

However, on closer inspection, RA(4) does not hold under this model. Under the causal model where C affects $Y^{c}(t=0)$, $Y^{c}(t=1)$, and A, conditioning on $Y^{c}(t=0)$ induces new dependence between Y^{θ} and A, and opens a backdoor path through exogenous variable $U_{Y(t=0)}$ and confounder C. This occurs because $Y^{c}(t=0)$ is a collider (two arrows go into the same variable). Conditioning on a collider opens a path that would otherwise be blocked.[92] This unblocked path, $A \leftarrow C - U_{Y(t=0)} \rightarrow Y^{\theta}$, is represented by the line between $U_{Y(t=0)}$ and C (labeled (ii) in figure 3.4). The path would be blocked if $Y^{c}(t=0)$ is not conditioned on.

Thus, to benefit from the potential to remove unmeasured confounding from the use of Y^{θ} as outcome, I need a new RA (9), which is not conditional on $Y^{c}(t=0)$:

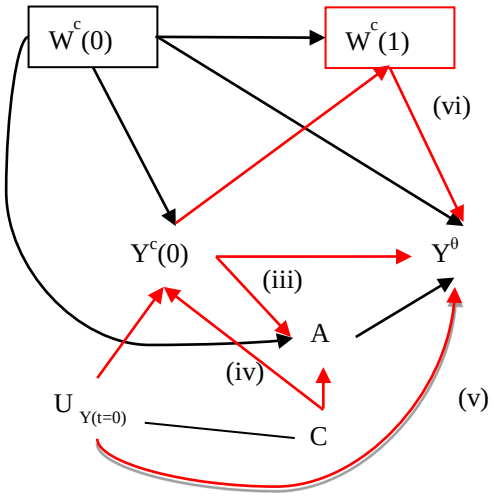
$$Y^{\theta}_{a} \perp A \mid W^{c}(t=0), W^{c}(t=1) \qquad (9)$$

It is important to note that I have arrived at the same conclusion with DAGs that others have reached using parametric equations and analysis of covariance. In the econometrics literature, the problem is recognized as the fact that the residual on Y^{θ} (in a parametric equation) is necessarily correlated with the lagged outcome, $Y^{c}(t=0)$, because both are a function of the random error on $Y^{c}(t=0)$ (i.e., a function of $U_{Y(t=0)}$ in my SCM).[99] Conditioning on $Y^{c}(t=0)$ has been demonstrated to bias the treatment effect estimate under this model where the errors on Y^{c} are serially correlated.[99] The method of differencing can still be applied if this correlation is thought to be negligible (i.e., possibly when the data are from a series of cross-sections of different individuals and/or the time between cross-sections is long).[100] However, RA(9) still

does not hold under this model without additional assumptions. I make these assumptions apparent with the use of the DAG shown in figure 3.5.

By not conditioning on $Y^c(t=0)$, I open up multiple new pathways from A to Y^θ: directly through $Y^c(t=0)$ ($A \leftarrow Y^c(t=0) \rightarrow Y^\theta$, labeled (iii) in figure 3.5); through C ($A \leftarrow C \rightarrow Y^c(t=0) \rightarrow Y^\theta$, labeled (iv)); and through $U_{Y(t=0)}$ ($A \leftarrow Y^c(t=0) \leftarrow U_{Y(t=0)} \rightarrow Y^\theta$, labeled (v)). Additionally, $W^c(t=1)$ is a descendant of collider $Y^c(t=0)$, and conditioning on $W^c(t=1)$ opens up the same pathway as conditioning on $Y^c(t=0)$ (i.e., $A \leftarrow C - U_{Y(t=0)} \rightarrow Y^\theta$). However, if I do not condition on $W^c(t=1)$, then I would open up new backdoor pathways through $W^c(t=1)$ (i.e., $A \leftarrow C \rightarrow Y^c(t=0) \rightarrow W^c(t=1) \rightarrow Y^\theta$ and $A \leftarrow Y^c(t=0) \rightarrow W^c(t=1) \rightarrow Y^\theta$ labeled (vi)).

Figure 3.5: Confounding by $Y^c(t=0)$

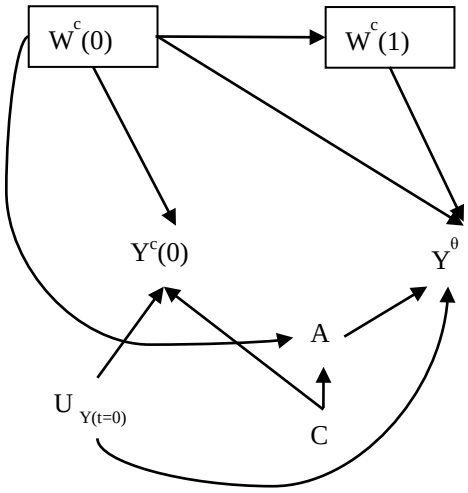


Therefore, I must be willing to make three additional exclusion restrictions for my casual parameter to be identifiable in a difference model: that $Y^c(t=0)$ must not affect A, $W^c(t=1)$ and $Y^c(t=1)$. The semi-parametric equation for Y^θ becomes:

$$Y^\theta = Y^c(t=1) - Y^c(t=0) = f_{Y(t=1)}(V, W^c(t=0), A, W^c(t=1), U_{Y(t=1)}) - f_{Y(t=0)}(V, W^c(t=0), U_{Y(t=0)})$$

where Y^θ is no longer a function of $Y^c(t=0)$ but is still a function of $U_{Y(t=0)}$ (see figure 3.6). Under this model, I can choose to either adjust for $W^c(t=1)$ or not (conditioning on $W^c(t=0)$ is sufficient and $W^c(t=1)$ is no longer a descendant of a collider).

Figure 3.6: Exclusion restrictions on $Y^c(t=0)$



In summary, RA(9) holds in the presence of unmeasured confounding from non-time varying factors, C, with a constant additive effect on $Y^c(t)$, only if $Y^c(t=0)$ does not affect A, $Y^c(t=1)$ and $W^c(t=1)$. The target causal parameter can now be identified as a new target parameter of the observed data distribution. The identifiability result applied to Y^θ becomes:

$E[Y^\theta_a \mid V, W^c(t=0), W^c(t=1)] =$
$E[Y^\theta_a \mid A=a, V, W^c(t=0), W^c(t=1)] =$
$E[Y^\theta \mid A=a, V, W^c(t=0), W^c(t=1)]$

where the first equality holds under the RA(9) and the second from the definition of the counterfactual outcome Y^{θ} under my new SCM (figure 3.6), giving me a new estimand for the ATE, Ψ^{II}:

$$\Psi^{II}(P_0) = E_{V,W(t=0),W(t=1)}[E[Y^{\theta} \mid A = 1, V, W^c(t = 0), W^c(t = 1)] - E[Y^{\theta} \mid A = 0, V, W^c(t = 0), W^c(t = 1)]] \quad (10)$$

which I refer to as the change score estimand.

Estimand III: Outcome $Y^c(t)$

Finally, there is an alternate difference-in-differences estimand that pools the outcome data from both time periods together. For this approach, I need to evaluate a third causal model for identifiability. Specifically, if I am willing to make additional assumptions on the underlying causal model such that:

$$E_{V,W(t=1),W(t=0)}[Y^c(t) \mid A=a, V, W^c(t=0), W^c(t=1)] = E_{V,W(t)}[Y^c(t) \mid A=a, V, W^c(t)], \text{ for } t = 0, 1 \quad (11)$$

then I have the following identifiability result under the new SCM:

$E[Y^{\theta}_a \mid V, W^c(t=0), W^c(t=1)] =$
$E[Y^{\theta}_a \mid A=a, V, W^c(t=0), W^c(t=1)] =$
$E[Y^c(t=1) \mid A=a, V, W^c(t=0), W^c(t=1)] - E[Y^c(t=0) \mid A=a, V, W^c(t=0), W^c\ t=1)] =$
$E[Y^c(t=1) \mid A=a, V, W^c(t=1)] - E[Y^c(t=0) \mid A=a, V, W^c(t=0)]$

As with estimand II, the first equality in the identifiability result holds under the RA(9). The last equality holds under assumption (11) (i.e., by substituting t=1 and t=0 for t), giving me a third estimand for the ATE:

$$\Psi^{III}(P_0) = E_{V,W(t)}[(E[Y^c(t = 1) \mid A = 1, V, W^c(t = 1)] - E[Y^c(t = 0) \mid A = 1, V, W^c(t = 0)]) - (E[Y^c(t = 1) \mid A = 0, V, W^c(t = 1)] - E[Y^c(t = 0) \mid A = 0, V, W^c(t = 0)])] \quad (12)$$

I refer to this final estimand as the pooled outcome estimand. However, there may be additional restrictions on the allowed data distribution for this identifiability result to hold. Starting with the SCM established for the change score estimand (Ψ^{II}), I work through the model separately at each time point. At time t=1, assumption (11) becomes:

$$E_{V,W(t=1),W(t=0)}[Y^c(t=1) \mid A=a, V, W^c(t=0), W^c(t=1)] = E_{V,W(1)}[Y^c(t=1) \mid A=a, V, W^c(t=1)],$$

which will hold if $Y^c(t=1)$ is independent of $W^c(t=0)$ given V, A, and $W^c(t=1)$. I can use the DAG shown in figure 3.7 to check whether my SCM implies this conditional independence. Under my current model, assumption (11) fails at t=1 because of two unblocked paths: the direct path from $W^c(t=0)$ to $Y^c(t=1)$ (label (vii) in figure 3.7); and the paths through collider A (i.e., $W^c(t=0) - C \rightarrow Y^c(t=1)$ label (viii) in figure 3.7). Therefore, for assumption (11) to hold at t=1, I need to add two new exclusion restrictions: that $W^c(t=0)$ does not affect $Y^c(t=1)$ and does not affect A (see figure 3.8).

Figure 3.7: Test of independence at t=1

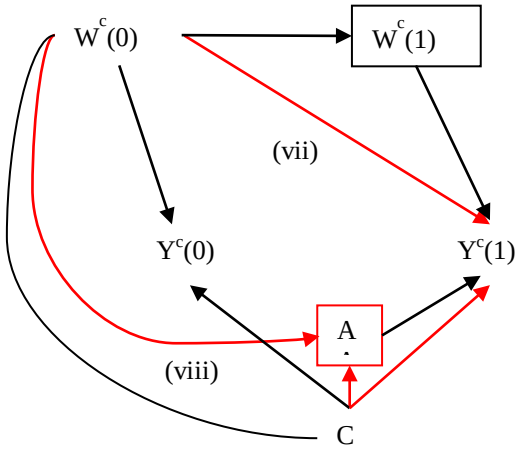
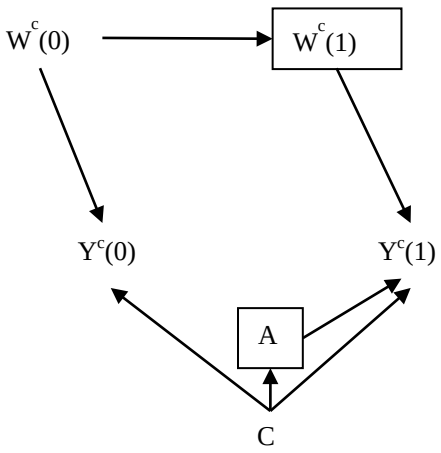


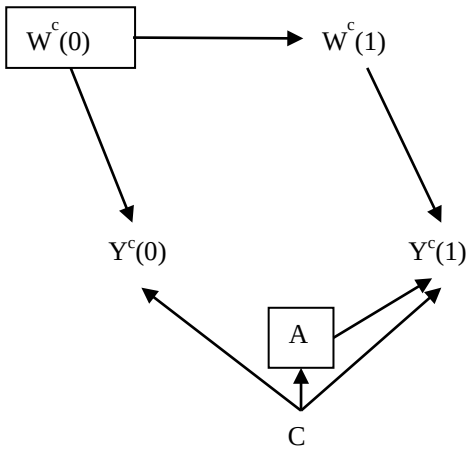
Figure 3.8: Exclusion restrictions on W(0)



Similarly, at time t=0, assumption (11) becomes:

$E_{V,W(t=1),W(t=0)}[Y^c(t=0) \mid A=a, V, W^c(t=0), W^c(t=1)] = E_{V,W(0)}[Y^c(t=0) \mid A=a, V, W^c(t=0)]$

Figure 3.9: Test of independence at t=0



and I verify with a DAG that my SCM implies $Y^c(t=0)$ is independent of $W^c(t=1)$ given V, A, and $W^c(t=0)$ (figure 3.9). No additional exclusion restrictions are required.

Note that I cannot add any arrows back that were removed for estimand II (i.e., $Y^c(t=0)$ cannot affect A, $W^c(t=1)$ or $Y^c(t=1)$. Under the additional restriction assumptions that $W^c(t=0)$ does not affect A and $Y^c(t=1)$, my causal target parameter, the ATE, is equivalent to estimand III. In settings where background knowledge makes it plausible to assume this more restrictive causal model, alternative estimation approaches offer some important advantages over traditional approaches, which will be discussed in chapter 4.

3.2.4 Illustration of Results Using Simulated Data

In this section, I present a series of simulations to demonstrate the need for the additional exclusion restrictions for the difference-in-differences estimands (Ψ^{II} and Ψ^{III}). The programming language R, version 2.13.1, was used for the simulations (the code is available in Appendix A2). As with the DAGs, I exclude the observed village factors, V, from the simulations. I present eight scenarios based on different SCMs represented by the DAGs in the previous section. In all cases, $Y^c(t)$, $W^c(t)$ and C are continuous, normally distributed and a

function of additive linear terms. Treatment variable, A, is dichotomous and the true parameter of interest, the ATE, has a value of 1. For each scenario and estimand, linear regression with main terms was used to estimate the relevant conditional expectation from a sample of 100,000 observations. These estimates are reported in table 3.2.

The first simulation is based on the starting SCM for the post treatment estimand (Ψ^{I}) represented in figure 3.1. Under this model, RA(1) holds. I obtain identical estimates of the target parameter whether the outcome is defined as $Y^{c}(t=1)$ or Y^{θ} (figure 3.2 and RA(4)). The estimate is nearly equal to the target parameter value of 1 (simulation #1, table 3.2). However, when I introduce an unmeasured confounder, C, in the second simulation, RA(1) and RA(4) no longer hold and the estimate diverges from the truth (simulation #2). This result is in keeping with a backdoor pathway being open from A to outcome $Y^{c}(t=1)$ through C (path (i) in figure 3.3) or with dependence between Y^{θ} and A through $U_{Y(t=0)}$ and confounder C (path (ii) in figure 3.4).

Table 3.2: Estimates for the ATE under various models and sample size (true value =1)

Sim. #	**Figure**	**Assumptions**	**Conditional on $Y^{c}(t=0)$**	**Estimate**	
1	3.1 & 3.2	RA(1) or (4), no unmeasured confounders, A does not affect $W^{c}(t=1)$	Yes	Ψ^{I}	0.99
2	3.3 & 3.4	RA(1) or (4) with unmeasured confounder C, A does not affect $W^{c}(t=1)$	Yes	Ψ^{I}	3.43
3	3.5	RA(9) with unmeasured confounder C, A does not affect $W^{c}(t=1)$, $Y^{c}(t=0)$ affects A, $W^{c}(t=1)$, and $Y^{c}(t=1)$	No	Ψ^{II}	4.78
4	3.6	RA(9) with unmeasured confounder C, A does not affect $W^{c}(t=1)$, and $Y^{c}(t=0)$ does not affect A, $W^{c}(t=1)$, or $Y^{c}(t=1)$	No	Ψ^{II}	0.98
5	N/A	Same as simulation 4 but $Y^{c}(t=0)$ affects A	No	Ψ^{II}	-1.78
6	3.7	RA(9) with unmeasured confounder C, A does not affect $W^{c}(t=1)$, $Y^{c}(t=0)$ does not affect A, $W^{c}(t=1)$, or $Y^{c}(t=1)$ Assumption (11) but $W^{c}(t=0)$ affects A and $Y^{c}(t=1)$	No	Ψ^{III}	6.64
7	3.8 & 3.9	RA(9) with unmeasured confounder C, A does not affect $W^{c}(t=1)$, $Y^{c}(t=0)$ does not affect $W^{c}(t=1)$, A, or $Y^{c}(t=1)$ Assumption (11), and $W^{c}(t=0)$ does not affect A or $Y^{c}(t=1)$	No	Ψ^{III}	1.02
8	N/A	Same as simulation 7 but $Y^{c}(t=0)$ affects A	No	Ψ^{III}	-0.99

Switching to my change score estimand (Ψ^{II}), I demonstrate that the estimate for the ATE diverges from 1 when not conditioning on $Y^{c}(t=0)$ (simulation #3) because it opens up new pathways from A to Y^{θ} (paths (iii) to (vi) in figure 3.5). By adding the necessary exclusion restrictions for estimand II in the fourth simulation (i.e., figure 3.6), the estimate once again nearly equals the target value (simulation #4). These results are comparable to those for the post

treatment estimand with no unmeasured confounding (simulation #1). However, in simulation #5, I add that $Y^c(t=0)$ affects A into the previous scenario for estimand II. In this fifth scenario, estimand II will diverge from the truth.

The sixth simulation represents the model for my pooled outcome estimand (Ψ^{III}), where at time t=1, $Y^c(t=1)$ is not independent of $W^c(t=0)$ given V, A, and $W^c(t=1)$ (figure 3.7). As expected, the estimate for estimand III diverges from 1 (simulation #6). However, when the paths from $W^c(t=0)$ to A and $Y^c(t=1)$ are removed (figure 3.8), estimand III is equal to the ATE (simulation #7). Finally, in simulation #8, I add that $Y^c(t=0)$ affects A into the previous scenario for estimand III, and the estimate once again diverges from the truth. As with simulation #5, this last simulation demonstrates that even if we can accept all the other exclusion restrictions for estimand III, we still must be willing to accept that $Y^c(t=0)$ does not affect A for the difference-in-differences estimands to equal the target parameter.

In summary, the above simulations show that when there is an unmeasured confounder, the post treatment estimand is not equal to the ATE whereas the change score and pooled outcome estimands might be, but only under additional assumptions. I demonstrate that even with an additive constant confounder C, I can get into trouble by using these latter two estimands (Ψ^{II} and Ψ^{III}) if $Y^c(t=0)$ affects A (i.e., $Y^c(t=0)$ is a confounder).

3.3 Discussion

Pre-post program evaluations (with data from treatment and control groups) present investigators with multiple causal models to choose from for identifying a causal effect of the program. Causal assumptions are necessary to obtain a valid estimate of a casual effect (e.g., the ATE), and each of these models relies on a different set of assumptions. However, the causal model (or SCM) needs to be defined before committing to a statistical model (as opposed to selecting an estimand based on the estimation procedure it allows). Specifically, the SCM incorporates expert knowledge about the data generating process that gave rise to the observed data. Any assumptions necessary to obtain a valid estimate of the desired causal effect should be reflected in the SCM and supported by this knowledge. The step of evaluating the assumptions for a given study should not be overlooked. It is up to the investigator to check these assumptions prior to selecting a model and proceeding with estimation. Failure to do so can result in choosing an estimand that is not equivalent to the target casual parameter. In this chapter, I use the structure of an existing program evaluation with pre-post data as the basis for defining several commonly used causal models. Through a series of graphical models (DAGs) and simulations, I explore the associated assumptions and identifiability result for three separate estimands. Most importantly, I demonstrate that the popular difference-in-differences model requires a considerable number of exclusion restrictions to express the target casual parameter as a parameter of the distribution of the observed data.

First I define the outcome as the post treatment value, $Y^c(t=1)$, and present a causal model that requires a minimal set of exclusion restriction assumptions for identification of the ATE (estimand I). Under the key assumption of no unmeasured confounding, the simple post treatment estimand (Ψ^I) equals my target parameter (the ATE) (simulation #1 in table 1). As expected, Ψ^I and the ATE diverge (i.e., are no longer equal) if an unmeasured factor, C, is introduced that confounds the relationship between treatment and outcome (simulation #2).

Since unmeasured cofounding is a realistic scenario in observational studies, it is not surprising that a difference-in-differences approach is often favored to try to address this issue. A differencing model is advantageous in that it "subtracts out" the effect of unmeasured confounders with a constant additive effect on the outcome at the two time points. The commonly accepted identifying assumption for the difference-in-differences estimand is a randomization assumption (RA(9)) typically referred to in the econometrics literature as the parallel trend assumption. However, through a step-by-step process of checking graphical models, I show that several exclusion restrictions are necessary for the difference-in-differences estimand to equal the ATE. In order to take advantage of this approach, I must be willing to assume that the lagged outcome, $Y^c(t=0)$, does not affect treatment, A, the post-treatment covariates, $W^c(t=1)$, or the post-treatment outcome, $Y^c(t=1)$. These are very strong assumptions about the lagged outcome! Under conditions where these restrictions do not hold, Ψ^{II} and Ψ^{III} will generally not be equivalent to the ATE (as illustrated with simulations #3, 5, 6, and 8). In fact, a difference-in-differences estimand has the potential to diverge further from the wished for causal effect than the post treatment estimand adjusting for all baseline covariates, even in the presence of an unmeasured confounder with a constant additive effect.

The exclusion restrictions become more numerous for the model that pools the outcome from both time periods (Ψ^{III}). In order for the pooled outcome estimand to be equal to the causal parameter of interest, I must add to the list of assumptions for the change score estimand (Ψ^{II}) that $W^c(t=0)$ does not affect A or $Y^c(t=1)$.

Although the exclusion restriction assumptions for difference-in –differences models may seem unrealistic, it is important to note that they are often applied to data from serial cross-sections of different persons from the same communities separated in time by many years.[100] It is possible under certain conditions that the pre-intervention outcome and covariates do not directly affect the post-intervention outcome and covariates, and are associated with post intervention outcome and covariates due only to fixed community level factors that affect both. In other words, $Y^c(t=0)$ may be predictive of $Y^c(t=1)$, but only due to shared common causes C or V. The advantage of controlling for unmeasured fixed effects with a difference-in-differences estimand must be weighed against what is known about the underlying data generating system and the associated model assumptions.

In this chapter, I demonstrate the power (and importance) of using graphical models (DAGs) to make the assumptions underlying a causal model transparent. The graphs prove to be invaluable tools for locating sources of dependencies among variables from confounders or colliders that may result in bias. Ideally, researchers should spend the time working with DAGs prior to conducting a study in order to collect the necessary data for a valid analysis. In an *ex-post facto* evaluation, it falls to the analyst to make use of DAGs, expert opinion, and other tools at their disposal, before proceeding with estimation. For example, given that the statistical model is semi-parametric, some of the exclusion restrictions are testable (i.e., that $Y^c(t=1)$ is independent of $W^c(t=0)$ given A, $W^c(t=1)$ and V).

In summary, my results reveal an important issue of identifiability that is not clearly articulated in the published literature. In the context of evaluating a community level intervention with pre-post data, I am confronted with a trade-off between statistical models that require expert

knowledge about the observed data before choosing one over the other. If my knowledge is sufficient to accurately represent the underlying data generating distribution, then my casual model may help me choose between estimands (e.g., whether the post treatment estimand is closer to the ATE than the pooled outcome estimand). In many cases, however, my knowledge will be insufficient and I won't know that the SCM holds for either estimand (or know which estimand is closer to my target parameter). Importantly, if I have strong evidence that a) there is important unmeasured confounding, and that b) the data do not support any other assumptions on which my identifiability results rely, then the target parameter is not identifiable. I cannot disregard this evidence; I risk getting a biased estimate. Instead, it is at this juncture that I must consider redefining my research question and target parameter before proceeding. The threat to validity from selecting a statistical model without understanding the underlying assumptions transcends my work and is applicable to any evaluation of an intervention.

Chapter 4: Effect Estimation – a Comparison of Methods

4.1 Introduction

In chapter 3, I illustrate the importance of using structural causal models and directed acyclic graphs (DAGs) for assessing identifiability prior to committing to a statistical model and estimand to target for estimation. Using a program evaluation from Madagascar with pre-post data, I identify three common statistical parameters which, under different assumptions, are equivalent to my causal target parameter of interest, the average treatment effect (ATE). I describe a trade-off between choosing a difference-in-differences estimand that remains equivalent to the ATE in the presence of a certain type of unmeasured confounder (i.e., the change score estimand or the pooled outcome estimand), and a model that assumes no such confounding exists (i.e., the post treatment estimand). The focus of this chapter is to compare estimates of the ATE of the Madagascar program on children's nutritional status using the observed data. Specifically, I apply three different estimation methods to each of the three estimands from chapter 3, and compare the resulting estimates and their confidence intervals. The first method, traditional parametric regression, is common across all disciplines. The second, inverse probability of treatment weights (IPTW or propensity score weighting), [64-66] has become popular in epidemiology but is also used by economists. And the third, a new method, has only recently been applied in epidemiology: targeted maximum likelihood estimation (TMLE).[58, 67] I briefly present each of the methods here and discuss them in more detail in the methods section.

Different estimators require estimators of different components (or parameters) of the observed data distribution. In my study, these components include the Q parameter, the treatment mechanism, and the empirical distribution of the covariates. The Q parameter (denoted as $\overline{Q_0}$) is the conditional mean of the outcome given treatment and covariates and is the basis of traditional parametric regression. The treatment mechanism (denoted as g_0) is the conditional distribution of treatment under the observed data distribution and is the key component for the IPTW estimator. TMLE makes use of both $\overline{Q_0}$ and g_0.

Estimation can be accomplished in a single step with parametric regression if the statistical model for the ATE is a linear equation without interaction (see section on parametric regression under estimation procedures). Covariate imbalance across treatment groups is addressed by conditioning on a set of measured confounders in the regression. A major limitation of traditional regression is that it relies on correctly specifying the full parametric model for $\overline{Q_0}$ in order to consistently estimate the target parameter of interest. However, I am unable to defend most of the parametric model assumptions imposed in traditional regression. For example, the inclusion of a continuous poverty index in a linear regression model for the effect of an intervention on nutritional status assumes that the relationship between nutritional status and poverty is linear and not quadratic or of some other functional form. The addition of higher-ordered terms or interaction terms with other variables, such as gender, may improve the model fit. But I don't know *a priori* what form these relationships should take, and the true relationship between nutritional status and poverty may be of a more complicated functional form than what I can approximate by these simple terms. Finally, as terms are added or removed from the model, the interpretation of the coefficient on the exposure changes. Traditional parametric regression provides an estimate of the ATE that will be biased if the model is incorrectly specified.

IPTW, on the other hand, is a two-step process that aims to address covariate imbalance by re-weighting the sample population into a balanced hypothetical population, or pseudo-population, in which the exposure is independent of the measured confounders (under my randomization and ETA assumptions).[65, 101] In the first step, the treatment mechanism, g_0, is estimated and used to calculate inverse probability of treatment weights (see section on IPTW under estimation procedures). In the second step, a weighted regression of the outcome is fit on treatment using a model that no longer includes the confounders. In this way, IPTW avoids specifying a parametric model for the expectation of the outcome given treatment and confounders, and by doing so, avoids the potential for introducing bias due to the model misspecification implied by this approach. However, IPTW relies on consistent estimation of the treatment mechanism, g_0, to obtain an unbiased estimate of the target parameter.

TMLE is multi-step process that involves estimation of both the Q parameter and the treatment mechanism in estimating the causal effect (see section on TMLE under estimation procedures). Importantly, TMLE implements a final bias reduction step to estimate the target parameter of interest with minimal bias. While the consistency of parametric regression relies wholly on consistent estimation of the conditional mean of Y given treatment and covariates, and the consistency of IPTW relies wholly on consistent estimation of the treatment mechanism, the consistency of TMLE relies on either consistent estimation of $\overline{Q_0}$ or consistent estimation of g_0. In this way, TMLE is considered a doubly robust estimator.

In addition, TMLE incorporates some of the best features of the other two estimation methods and avoids some of their problem areas. For example, TMLE shares an advantage with parametric regression in that they both belong to a class of estimators known as substitution estimators. As discussed in the introductory chapter, an estimator is a mapping function that takes as input an estimate of the distribution (where the estimated distribution is not necessarily an element of the statistical model) and returns as output an estimate of the true target parameter value (e.g., the ATE). A substitution estimator is a type of estimator that applies the same mapping (or function) that defines the target parameter to an estimate of the distribution of the data (where the estimated distribution is an element of the statistical model). Substitution estimators have the important advantage that they respect the constraints, or knowledge, incorporated in the statistical model.

Both parametric regression and TMLE are substitution estimators in that they apply the target parameter mapping, Ψ (e.g., estimands I to III), to an estimate of $\overline{Q_0}$ that respects the statistical model, in order to estimate the parameter of interest. Put more simply, but less precisely: these two substitution estimators plug an estimate of $\overline{Q_0}$ into the same function that defines Ψ. By comparison, IPTW is not a substitution estimator. The IPTW estimator applies a different mapping that is based on estimating a different part of the data distribution (i.e., g_0 that is not part of the parameter definition) and plugging this estimate into a different function (see section on IPTW under estimation procedures).

In addition, both parametric regression and TMLE are able to extrapolate to areas beyond the support of the data (i.e., to levels of high or low poverty that were not actually observed for a

given treatment assignment) when estimating the ATE. This can be an advantage in sparse data situations, but only to the extent that I am willing to assume a weak positivity assumption (i.e., that it is reasonable to extrapolate to these areas, see discussion of the ETA in chapter 3) and the statistical model is correctly specified.

Finally, both IPTW and TMLE are semi-parametric estimators. Their implementation requires specifying estimators for $\overline{Q_0}$ and/or g_0. Although parametric regression can be used to estimate g or Q, doing so would introduce parametric model specifications into methods that are otherwise non- or semi-parametric. Therefore, in order to minimize model misspecification, I incorporate a non-parametric, machine-learning tool, or SuperLearner (SL),[102] for the prediction steps in IPTW and TMLE. Briefly, SuperLearner is a data-adaptive tool that "learns" from the observed data by using a candidate set of algorithms (or estimators) and a pre-specified loss function that assigns a measure of performance to each of the algorithms. I provide a short description of SuperLearner in the Appendix (A3), but for more detail, I refer the interested reader to the Super Learning chapter of the van der Laan and Rose book on Targeted Learning.[58, 102]

The non-random assignment of treatment in Madagascar (or in any observational study) makes inferences about the program's effect susceptible to confounding bias if the comparison group is not exchangeable with the treated group on key predictors of the outcome. Therefore, it is critical to observe the key confounders and to optimally control for them in the estimation process. Each of the three methods discussed above offers a different approach to obtaining an estimate (that may or may not be biased) of the target causal parameter, given measured confounders. On the other hand, the exclusion of strong, unmeasured confounders from the statistical model will result in the divergence of the estimand (i.e., the statistical parameter) from the target causal parameter, regardless of the method of estimation. For example, using simulations in chapter 3, I demonstrate that the estimate of the ATE diverges from the true value when the necessary assumptions fail to hold for an estimand. However, the magnitude and direction of this divergence for the observed data from Madagascar is unknowable. For the sake of exploring the different identifiability solutions with the Madagascar data, I proceed with estimation of each of the three estimands defined previously in chapter 3.

4.2 Methods

4.2.1 Data

In 1999, the Government of Madagascar rolled out a national, community-based growth monitoring and nutrition education program (*SEECALINE*).[43] A nationally representative anthropometrics survey was performed in 420 communities in 1997/98, prior to the implementation of the program. The survey was administered to a random sample of 14,148 households, 12,814 of which had children five years of age and under. Both weight and length/height were obtained for the children, but only the weight data are used in my analysis.

An initial treatment assignment in 1999 was made at the community-level based on district-level prevalence of moderate underweight among children under five (moderate underweight is defined as weight-for-age z-score 2 standard deviations below the median of a reference population). The program was phased in through 2002, and expanded to include new communities impacted by severe weather conditions in 2000 or impacted by political instability

in 2002. In total, 3,600 project sites were reached. In 2004 a second nationally representative anthropometric survey was administered to 10,704 households, 9,296 with children under 5 years of age, in 446 program participating and non-participating communities (420 communities from the 1997/98 survey plus 26 new communities).

Table 4.1: Variable Description

Abbreviation	Description
V	Vector of time invariant community level covariates: † • Rural vs. urban location • Province • Population size • Presence of health facility • Road access (paved, unpaved, or none in wet season) • Water access (in any season) • Indicators for weather shocks in commune between '99 and '01
$W^c(t)$	Vector of individual level covariates, $W_i(t)$, (i=1,…,N), for each of the N individuals sampled in the community at time t=0,1, aggregated to the community level as a mean or proportion: • Maternal education (proportion with no education, proportion with primary only education) • Child age (mean age and proportion older than 1 year) • Child gender (proportion female) • Child birth order (mean rank)
A	SEECALINE administered at the community level
$Y^c(t)$	Community mean of individual level outcomes Yi(t) (i=1,…,N) for each of the N individuals sampled in the community at time t=0,1: • Weight-for-age z-score‡ of children 6-59 months
$U_V,…U_{Y(t)}$	Random variation for each endogenous variable that might include: • Characteristics of leadership in accepting the program (U_A) • Dispersion of the community across large distances (U_V) • Lack of a secondary school in the community (U_W) • Sampling procedure problems ($U_{Y(t)}$)

†Population size, presence of a health facility, and access by road or water information is only available at the community level in 2004. For the purposes of this paper, I assume that these factors did not change significantly from 1997.

‡A weight-for-age z-score value of +1 (-1) is equivalent to 1 standard deviation (SD) above (below) the median weight of the WHO reference population of well- nourished and healthy children of the same age and gender.[103]

I restricted the analytic sample to 410 out of the 446 communities. Twenty-six communities were excluded that were not part of the baseline survey in 1997, but were added later in 2004. I excluded another 10 communities (including the provincial capitals) from 6 urban districts because sites in these districts were opened in 2002 in response to a political crisis, such that the nature and the socio-economic context of the intervention differed substantially from the

remainder of the country. Descriptions of the observed variables are shown in table 4.1. The U's are unmeasured, exogenous variables.

4.2.2 Identifiability in the Madagascar Context

Prior to proceeding with estimation, I review whether the randomization assumption (RA) and experimental treatment assignment (ETA) assumption are likely to hold in the context of the observed data from Madagascar. Given that the statistical models for estimands I, II, and III are semi-parametric, some of the exclusion restrictions are testable and findings for these are presented in the results section (i.e., the independence assumptions for estimand III).

Randomization Assumption

Treatment assignment of the nutrition program was not random. Therefore, it is essential that I identify the factors that may differ systematically between treated and untreated communities. These factors (if observed) can then be included in the subset of conditioning variables necessary for the randomization assumption to hold. Forty-six out of 111 districts in Madagascar were targeted for intervention based on having a district-level prevalence of moderate underweight above the national average (43%) in 1997/98. Another 11 districts were included in the initial roll out because they had experienced drought or cyclones in 2000. Lack of district participation in 2004 is nearly perfectly predicted by these two indicators: most (92%) of communities in the non-targeted districts did not take up the program. However, there is more heterogeneity in participation status in the targeted districts: about 66% of communities in these districts in my sample took up the program by 2004. Other factors that influenced the actual roll out and implementation program implementation included the goal of achieving a coverage rate of 50% of children under 3 years of age by 2002 (i.e., targeting larger communities), and the requirement that the community be accessible by local transportation for most of the year for field supervision (i.e., by auto/motorbike, horse cart, or canoe). The presence of local non-governmental organizations was necessary to manage the field supervision. Inclusion of these factors in the conditioning set of variables (for the RA assumption to hold) may result in ETA assumption violations, which I explore in more detail later.

Observed variables that encompass the district and other selection criteria, and that are also expected to predict the final community-level outcome, include the mean community weight-for-age z-score (WAZ) in 1997 (or $Y^c(t=0)$), the population size of the community, indicators for road and water access in both the wet and dry seasons, and an indicator at the commune level of the occurrence of a drought or cyclone prior to 2002. I use the community mean WAZ at baseline (the lagged outcome) instead of the prevalence of moderate underweight in the community because the prevalence is simply a re-categorization of the outcome, and the two are very highly correlated (Pearson's $\rho = -0.91$). Additional variables included in V and $W^c(t=0)$ represent other factors that are generally considered associated with poor nutritional outcomes (i.e., low maternal education and lack of a hospital) and may be associated with treatment (see table 4.1).

Examples of unmeasured, exogenous factors that may cause random variation among the observed variables are listed in table 4.1. Of these, it is plausible that dispersion of the community across large distances (U_V) may act as an unmeasured confounder. Wide dispersion could influence the sample selection ($U_{Y(t)}$), and thus the nutritional status of the children included in the sample, as well as community cohesion, and thus the leadership choice to accept

the program into the community (U_A). In addition, very remote communities may have lacked a local non-governmental organization (i.e., could not receive treatment) and this remoteness might be reflected in nutritional outcomes.

Experimental Treatment Assignment

In addition to the RA, for the target statistical parameter to be defined in the structural causal model (SCM), each community must have some positive probability of both being treated and not being treated (experimental treatment assignment or ETA assumption). The ETA assumption (also known as the positivity assumption) states that for a binary treatment, A, the conditional probability of treatment given covariates, is bounded away from 0 and 1. (Note that in a true randomized control trial, the probability of treatment is independent of covariates and bounded away from 0 and 1 by design). If the ETA assumption is theoretically violated (i.e., the treatment is not possible or inevitable for certain values of the covariates), then the causal parameter is not identifiable without additional assumptions. Practical violations of the ETA may also occur if the true probability distribution is greater than zero, but, by chance, the probability equals or approaches zero in the available sample. In this case, while the causal parameter is formally identifiable, it may be poorly supported by the available data. Positivity violations can result in a biased estimate of the causal effect, regardless of the estimator, so it is important to assess the evidence in the context of my study.[96]

In the Madagascar example, the ETA assumption was not theoretically violated. Communities in non-targeted districts participated and communities in targeted districts did not. In addition, many non-participating communities received the program after 2004 as the program expanded. However, due to the fact that treatment was targeted to districts with the highest prevalence of malnutrition, that the sample is finite, and that the covariate data are high dimensional (some are continuous or multilevel), it is possible that the ETA assumption is practically violated. A formal diagnostic based on the parametric bootstrap is available for estimating the presence and magnitude of bias from positivity violations and near-violations.[104] However, in this chapter, I pursue several informal methods for investigating practical positivity violations.[96, 101]

Based on *a priori* knowledge of how the program was rolled out, I verify that both treated and untreated communities are represented in specific subsets of the data with low and high expected probability of treatment (i.e., in villages with low and high levels of malnutrition and in villages with and without road access). Due to the impossibility of checking every level of the covariates, I also examine the distribution of estimated probabilities of treatment given my covariates (also known as the propensity score). With respect to positivity violations, I verify that the estimated probabilities of treatment are bounded between 0.025 and 0.975 in my sample for each of the estimands (these bounds are the default for truncation in the TMLE package, see TMLE section for more detail). Unfortunately, neither of these checks quantifies the degree to which violations or near-violations threaten the validity of my causal effect estimate. However, evidence of heterogeneity in treatment within strata of the confounders can give me some confidence that the ETA assumption is reasonably held. Additional informal checks for positivity violations are discussed in the estimation section.

ETA violations may be addressed in a number of ways. Methods used prior to estimation include: 1) restricting the sample by trimming (or dropping) communities that have positivity

violations; 2) redefining the causal effect of interest to one that does not result in positivity violations (e.g., an average treatment effect among the treated (ATT) may meet this criterion); and 3) restricting the covariate adjustment set such that covariates responsible for positivity violations are excluded (must assume that these covariates are not strong confounders, although if they are weak confounders, I lose identifiability).[96] All three approaches change the parameter being estimated and are not implemented here. Methods that are applied in the estimation process include extrapolation based on subgroups with sufficient experimentation (which changes the semi-parametric statistical model by imposing additional assumptions), or truncation of extreme probabilities to some fixed values (which changes the target causal parameter). These latter two methods are discussed in the section on estimation procedures.

Regression to the Mean Bias

It is important to note that $Y^c(t)$ at both time points is estimated with sampling error and the inclusion of $Y^c(t=0)$ to control for confounding in estimands II and III can introduce a separate source of bias (i.e., distinct from bias due to conditioning on a collider). Specifically, if the outcome is measured with error, then inclusion of the lagged outcome, $Y^c(t=0)$, in a change score estimator leads to regression to the mean (RTM) bias.[105, 106] A common example of RTM is found in the clinical trials literature when a change in outcome pre and post treatment is of interest (i.e., for blood pressure or serum HDL cholesterol).[106, 107] Methods have been developed to correct for RTM.[106, 107] However, a good estimate of the reliability of the outcome measure is needed, either from repeated measures (within a short time frame) or a subset analysis using a gold standard instrument. Since I do not condition on $Y^c(t=0)$ in estimands II and III, corrections for the RTM are not implemented.

4.2.3 Estimation Procedures

In this section, I describe three estimation methods for each of the statistical target parameters: traditional parametric regression, inverse probability of treatment weights (IPTW),[64] and targeted maximum likelihood estimation (TMLE).[58, 67] As discussed, the different estimators require estimators of different components of the observed data distribution (i.e., g_0 and/or $\overline{Q_0}$), which vary for my three estimands.

I define the Q parameter for estimands I, II, and III, respectively, as:

For Ψ^I: $\overline{Q_0} = E_0 [Y^c(t=1) \mid A=a, V, W^c(t=0), Y^c(t=0), W^c(t=1)]$,

For Ψ^{II}: $\overline{Q_0} = E_0 [Y^\theta \mid A=a, V, W^c(t=0), W^c(t=1)]$, and

For Ψ^{III}: $\overline{Q_0} = E_0 [Y^c(t) \mid A=a, V, W^c(t), t]$

where $\overline{Q_0}$ represents the conditional mean of Y given treatment and covariates. The three variations on g are:

For Ψ^I: $g_0(a \mid V,W^c(t=0),Y^c(t=0),W^c(t=1)) = P_0(A=a \mid V,W^c(t=0),Y^c(t=0),W^c(t=1))$

For Ψ^{II}: $g_0(a \mid V,W^c(t=0),W^c(t=1)) = P_0(A=a \mid V, W^c(t=0), W^c(t=1))$

For Ψ^{III}: $g_0(a \mid V,W^c(t),t) = P_0(A=a \mid V, W^c(t), t)$

where g_0 represents the treatment mechanism applied to the true data generating distribution, P_0.

For ease of comparison, the community is the unit of analysis for all three estimators. The data are from two cross-sectional surveys with different individuals included in each year (i.e., Y(t) and W(t)) were measured on one set of subjects at time t = 0 and on another set of subjects at time t = 1). Selecting the community as the unit of analysis would be a limitation for repeated cross-sectional studies with few sites, but is feasible in my study of 410 communities.

A) Parametric Regression:
For the traditional parametric regression estimator, I impose the following generalized linear forms to the estimators for the post treatment and change score estimands (Ψ^I and Ψ^{II}), respectively:

$$E(Y^c(t=1) \mid V, A, W^c(t=0), Y^c(t=0), W^c(t=1)) = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 W^c(t=0) + \beta_4 Y^c(t=0) + \beta_5 W^c(t=1) \quad (4.1)$$

and

$$E(Y^\theta \mid V, A, W^c(t=0), W^c(t=1)) = \beta_0 + \beta_1 A + \beta_2 V + \beta_3 W^c(t=0) + \beta_4 W^c(t=1) \quad (4.2)$$

If these parametric statistical models are correctly specified (i.e., I was confident that the true data generating distribution fell in the family of distributions described by the models) then the coefficient β_1 on A is equivalent to my target parameter (the ATE). The ATE can then be estimated efficiently using maximum likelihood estimation (MLE). However, since my knowledge is inadequate to support such a model (i.e., I don't know the underlying functional form that describes how the mean outcome varies as a function of the covariates and the treatment), the coefficient β_1 is no longer necessarily equivalent to my target statistical parameter. In this case, an estimate of β_1 will generally give a biased estimate of my causal effect of interest. I include the post intervention covariates, $W^c(t=1)$, that I assume are not affected by A. Their inclusion may improve the precision in the estimator, and I test that the exclusion of $W^c(t=1)$ does not change my findings.

Note that although the coefficient β_1 describes a conditional association, it is equivalent to a marginal association in models (4.1) and (4.2) because these models assume that the conditional mean of Y given treatment and covariates is a linear function of A, and A does not interact with the covariates. A common misconception of traditional parametric regression estimators is that they confine us to a conditional association. For example, under a statistical model where A interacts with covariates, the conditional association is not equivalent to an average treatment effect (ATE). However, a marginal association can be obtained by adding a second stage to the estimation process. Specifically, the differences in mean outcome under treatment and no treatment are averaged over the empirical distribution of the covariates (e.g., $W^c(t)$ and V). This two-stage approach to estimating the ATE is comparable to standardization, where the marginal distribution of the covariates is used as the standard. It is also referred to as the g-computation estimator or the ML-based substitution estimator.[58]

For the third estimand (Ψ^{III}), I can choose to estimate $E_{V,W(t)}(E(Y^c(t) \mid A=a, V, W^c(t))$ separately for each time point, t, or alternatively, I can choose to implement a single estimator for $E(Y^c(t) \mid t, V, A, W^c(t)))$, pooling over time. I choose the latter. In addition, I have the option of implementing an individual-level analysis despite the serial cross-sectional nature of the data (i.e., different individuals are sampled at the two time points). This would take advantage of the knowledge that W(t) and Y(t) are drawn from the same individual for a given time t and possibly improve precision (i.e., if the paired data is more predictive of Y(t)). However, for comparability with the other estimands, I perform a community-level analysis for estimand III.

I impose the following functional form to the estimator of Ψ^{III}:

$$E(Y^c(t) \mid t, V, A, W^c(t)) = \beta_0 + \beta_1 A + \beta_2 t + \beta_3 (A*t) + \beta_4 V + \beta_5 W^c(t) \quad (4.3)$$

Once again, if I modify my semi-parametric statistical model for estimand III to assume that this parametric form is true, then the coefficient β_3 on the interaction term is equivalent to the target statistical parameter, Ψ^{III}, for the average treatment effect (ATE). Again, however, in the semi-parametric statistical model implied by the SCM, this is not necessarily true, and β_3 will be an unbiased estimate of Ψ^{III} only if the parametric model is correctly specified. The term "difference-in-differences" stems from the fact that the ATE estimate in this approach is the mathematical equivalent of the difference in the control arm over time (β_2) subtracted from the difference in the treatment group over time ($\beta_2 + \beta_3$), or the difference-in-differences of the treatment vs. control group over time (β_3). In keeping with the methods used for the first two estimands, I implement a generalized linear model with MLE for this estimator.

B) IPTW

As mentioned in the introduction, IPTW re-weights the sample population into a balanced pseudo-population. Each community is given a weight that is inversely proportional to an estimate of the community's probability of having received its observed treatment status, conditional on its measured confounders (obtained by estimating g_0 and then generating predicted values for the communities' observed treatment status). In other words, an IPTW estimator maps the empirical data distribution to a parameter estimate, beta (β), based on an estimator of the treatment mechanism.

For the first estimand, Ψ^I, I implement an IPTW estimator for β in the following saturated model:

$$E_{(V,W^c(t=0),Y^c(t=0),W^c(t=1)\mid A)} E[Y^c(t=1) \mid A, V, W^c(t=0), Y^c(t=0), W^c(t=1)] = \beta_0 + \beta_1 A \quad (4.4)$$

The estimate for the treatment effect, or the coefficient β_1, is obtained by fitting a weighted regression of $Y^c(t=1)$ on A according to the model (4.4). The weights are community specific and equal to the inverse of the predicted probability of treatment received, conditional on the covariates. Specifically for Ψ^I:

$$\text{weight}^I = 1 / g_n(A=a \mid V, W^c(t=0), Y^c(t=0), W^c(t=1)) \quad (4.5)$$

where g_n is the estimate of g_0. A numerator of one in equation (4.5) implies the use of unstabilized weights. In the presence of sparse data and extreme weights, the numerator can be

changed in such a way as to stabilize the weights (i.e., give greater weight to areas where the joint distribution of treatment and covariate(s) are well supported), but this approach applies when the model is unsaturated.[64, 96] In a fully saturated model (e.g., 4.4), changing the numerator won't make a difference, and so I employ a numerator of one. As with the parametric regression based estimator, I include the post intervention covariates, $W^c(t=1)$, in the denominator to potentially improve the precision of the estimator.

Similarly, for Ψ^{II}, I implement an IPTW estimator for β in the following model:

$$E_{(V,W^c(t=0),W^c(t=1)|A)}E[Y^\theta| A, V, W^c(t=0), W^c(t=1)] = \beta_0 + \beta_1 A \quad (4.6)$$

and estimate the coefficient β_1 by fitting a weighted regression of Y^θ according to model (4.6). The weights are equal to the inverse of the predicted probability of treatment received, conditional on the covariates V, $W^c(t=0)$ and $W^c(t=1)$, excluding $Y^c(t=0)$:

$$\text{weight}^{II} = 1 / g_n(A=a | V, W^c(t=0), W^c(t=1)) \quad (4.7)$$

Finally for Ψ^{III}, I implement an IPTW estimator for β, pooling over time points, for the following saturated model:

$$E_{(V,W^c(t)|A,t)}E[Y^c(t)| t, A, V, W^c(t)] = \beta_0 + \beta_1 A + \beta_2 t + \beta_3(A*t) \quad (4.8)$$

The estimate for the treatment effect is now the coefficient β_3 on the interaction term, which I obtain by fitting a weighted regression of $Y^c(t)$ according to model (4.8). Note that the weights for estimand III vary for $W^c(t)$ at the two time points, and exclude $Y^c(t=0)$:

$$\text{weight}^{III} = 1 / g_n(A=a | V, W^c(t),t). \quad (4.9)$$

The consistency of the IPTW estimator relies on consistent estimation of the treatment mechanism (g_0) used to calculate the weights in 4.5, 4.7, and 4.9 above. It is common practice to revert to a parametric statistical model for estimating g_0, but this approach introduces the potential for model misspecification (and consequently bias) that I am trying to avoid. Therefore, I use SuperLearner for the estimation of g_0, which respects my non-parametric statistical models.

In addition, the IPTW estimator is particularly susceptible to bias arising from violations of the ETA assumption. As discussed previously, IPTW is not a substitution estimator, and as such cannot extrapolate to areas with zero experimentation within certain levels of the covariates (true ETA violation). Under these conditions, the IPTW estimator is biased. In areas with minimal experimentation (near violation) a few communities may receive extremely large weights. Extreme weights can lead to finite sample bias and increased variance, although these weights can be truncated to some fixed minimum and maximum values.[101] Truncation can reduce the estimator variance, but the process introduces a new form of bias (i.e., the estimator of g is no longer consistent), and depending on the setting, it may or may not increase the mean square error (MSE)[96, 104] Extreme weights were not generated for any of the three estimands in my original sample. However, highly improbable observations occurred in the bootstrap samples

used for inference (see inference section), such that estimates of g_0 were bounded between 0.025 and 0.975 for the bootstrap estimates.

C) TMLE

TMLE is a doubly robust estimator that is designed to reduce bias in the estimator of the causal parameter of interest. The following description of the method applies to all 3 estimands.

TMLE is a multi-step process. The first step is to obtain an initial estimate of $\overline{Q_0}$, which can be performed with a parametric regression estimator as described previously. However, as with my implementation of IPTW, I avoid imposing unnecessary (and unrealistic) parametric model assumptions with TMLE. Specifically, I implement SuperLearner for estimating the Q parameter and average the differences in mean outcomes under treatment and no treatment over the empirical distribution of covariates (as I would do to apply the g-computation estimator).

TMLE takes the estimation process one step further. The SuperLearner estimator is designed to obtain the best predictive fit for the full conditional distribution of the outcome, $\overline{Q_0}$, based on an optimal bias-variance tradeoff procedure. However, I am not interested in the full conditional distribution of the outcome (or even its conditional mean), but rather in the mean difference in potential outcomes under treatment and no treatment. In other words, the optimization in SuperLearner does not give me the optimal bias-variance tradeoff for my actual parameter of interest. Therefore, an estimate of my target parameter from the initial fit of $\overline{Q_0}$ will be overly biased. In order to resolve this discrepancy, TMLE implements a targeting step to reduce the bias for the statistical parameter of interest (equal under assumptions to my causal parameter). The initial estimator of $\overline{Q_0}$ is updated in a fluctuation procedure using a "clever covariate," which is a function of the treatment mechanism, g_0.[58] Once again, I implement SuperLearner to obtain a consistent estimator of the treatment mechanism. The updated estimates of the predicted values of the outcome under each treatment condition are then used to obtain the target parameter of interest, the ATE. In this way, TMLE removes all asymptotic residual bias of the initial estimator for the target parameter, as long as I have a consistent estimator for g_0 or $\overline{Q_0}$.[58] As with IPTW, estimates of g_0 are bounded between 0.025 and 0.975 for the bootstrap samples used for inference.

In summary, TMLE has multiple advantages. As with IPTW, I am able to retain the semi-parametric nature of TMLE by the inclusion of SuperLearner in the prediction steps. As with parametric regression, TMLE is a substitution estimator, and as such, can extrapolate into areas not supported by the data within the bounds of the statistical model. Both parametric regression and TMLE are less sensitive to outliers and sparse data than IPTW in this regard, if extrapolating is justified and the statistical model is correctly specified. Finally, TMLE is advantageous over both of the other methods in that it is targeted to the parameter of interest and is "doubly robust." In estimating the causal effect of interest, TMLE is consistent if either the estimate of $\overline{Q_0}$ or the estimate of g_0 is consistent. If both are consistent, then the TMLE estimator of the target parameter is efficient. However, if the positivity assumption fails to hold, then TMLE relies entirely on consistent estimation of $\overline{Q_0}$.

4.2.4 Statistical Inference

Standard errors were estimated using a non-parametric bootstrap with 200 replications. For each bootstrap sample, 410 communities were sampled with replacement and all nine estimates obtained before drawing the next sample. The confidence intervals were calculated both parametrically (assumes a normal distribution for the estimator) and non-parametrically by ordering the bootstrap estimates and taking the 2.5^{th} and 97.5^{th} percentile values. I also present robust influence curve based confidence intervals using standard packages in R and the TMLE package.

4.2.5 Software Packages

Software package Stata/MP 10.1 was used for building the datasets and the programming language R, version 2.13.1, was used for all of the analyses. Analyses used the following publically available R packages: Super Learner Prediction version 1.1-18 and TMLE: Targeted Maximum Likelihood Estimation of Point Treatment Effects, version 1.1.1. Both packages are available on the Comprehensive R Archive Network (CRAN). Candidate SuperLearner algorithms that were included for predicting g were generalized linear models, Bayesian linear models, elastic nets, generalized additive models, step-wise regression, k-nearest neighbors, and neural networks. Candidate algorithms that were included for predicting Q were generalized linear models, Bayesian linear models, elastic nets, generalized additive models, step-wise regression, and polynomial spline regression. An extension of the publically available TMLE package, obtained directly from its author, Dr. Susan Gruber, was used to implement the TMLE estimator for estimand III (now available as an updated TMLE package, version 1.2, on CRAN).

4.3 Results

4.3.1 Exclusion restriction assumptions

All Estimands: A does not affect $W^c(t=1)$

For all estimands, I make the assumption that A does not affect $W^c(t=1)$. This is a causal assumption in that it expresses what would happen to $W^c(t=1)$ under changing conditions of treatment (e.g., setting A equal to 0 or 1). Although I cannot test a causal assumption directly with a statistical test, there are implications of causal assumptions that are testable. [93] For example, I find that A is not significantly associated with any of the variables in $W^c(t=1)$, when controlling for the corresponding variable in $W^c(t=0)$ using a series of parametric regressions. In addition, I discuss the sensitivity to removing $W^c(t=1)$ form the set of conditioning variables for estimands I and II in the estimation results section.

Estimands II & III: $Y^c(t=0)$ does not affect A

For the change score and pooled outcome estimands (Ψ^{II} and Ψ^{III}), I need to make the assumption that $Y^c(t=0)$ does not affect A. Selection into the treatment group was based (at least in part) on prevalence of moderate underweight in the community (aggregated up to the district level), making it difficult to accept this assumption. In a traditional logistic regression to predict A, the model that includes only $Y^c(t=0)$ explains approximately 7% of the variance in A. A model that includes only V, $W^c(t=0)$ and $W^c(t=1)$ explains about 15% of the variance. However, this increases to 22% with the addition of $Y^c(t=0)$ (likelihood ratio test is statistically significant, p-value < 0.001). I explore the implications of adding that $Y^c(t=0)$ affects A back into the models for estimands II and III in the discussion section.

Estimand III: Tests of independence

The additional assumptions required for the pooled outcome estimand (Ψ^{III}) can be tested. Based on parametric regression, none of the variables in $W^c(t=1)$ were found to be significantly associated with $Y^c(t=0)$ given V, A, and $W^c(t=0)$ at the 5% level. However, one of the six variables in $W^c(t=0)$ has a statistically significant association with $Y^c(t=1)$ given V, A, and $W^c(t=1)$. Specifically, mean infant birth rank in 1997 is negatively associated with mean weight-for-age in 2004 (p-value of 0.03). The magnitude and significance of assocation do not change substantially with the inclusion of $Y^c(t=0)$ and/or the exclusion of A from the regression. In very low income settings, children born first (rank = 1) often have worse nutritional outcomes than their later born siblings (the older children stop receiving breast milk when their younger siblings are born). However, at the community level, the mean birth rank in the sample is more likely to be reflective of family size, in which case I would expect a negative correlation of increasing family size with worse nutritional outcomes. In fact, there is a statistically significant negative correlation of mean infant rank and mean WAZ in both of the survey years in Madagascar (as well as in a recent 2011 survey). Therefore, it is possible that family size in 1997 is predictive of mean WAZ in 2004, and that this association is not blocked by V, A, or $W^c(t=1)$. It is also possible that the association is due to chance.

4.3.2 Checks for Positivity Violations

The treated communities have on average higher prevalence of underweight, as expected (39% vs. 30%). The minimum prevalence of underweight among the treated communities was about 5%, and the maximum, 95% (median 38%). The minimum prevalence of underweight among the untreated communities was 0%, and the maximum, 70% (median 29%). Approximately 40% of the treated villages had a prevalence of underweight greater than 43% in 1997.

Since provincial capitals were excluded from the sample and urban communities are generally better off economically than rural ones, I tabulated province and urban location among villages with high prevalence of underweight (above 43%). There are only 12 communities in the sample that are urban with a high prevalence of underweight, and there are only 1-3 of these in each of the 6 provinces. As a result, 4 of the provinces have empty cells for urban communities with high prevalence of underweight when cross-tabulated by treatment. In addition, of the 8 urban communities from the capital province of Antananarivo, none had received treatment in 2004, regardless of underweight prevalence (although 4 later received treatment). Therefore, there is evidence that the ETA assumption is violated for certain types of communities. This violation will lead to bias in the IPTW estimator and require the parametric regression and TMLE to extrapolate to these areas lacking in experimentation.

Finally, I checked for variability in treatment status by province and history of cyclone, as well as by road or water access to the communities. The villages in the northern-most province of Antsiranana are nearly dichotomized into treatment groups by whether or not they experienced a cyclone pre-2002. However, there were no empty cells (one treated community without cyclones and 2 non-treated communities that did experience a cyclone). There is reasonable heterogeneity by treatment status by transportation access in all provinces.

Despite the evidence of possible positivity violations in the above descriptive statistics, the predicted probability of treatment given covariates falls within the bounds of 0.025 and 0.975 for all communities and all three estimands (see min/max values in table 4.2). As a consequence,

there are no extreme weights. However, the absence of extreme weights is not definitive proof of positivity (as is evidenced by the ETA violation found with the cross-tabulated data). The range of probabilities in the untreated group is comparable to that of the treated group (table 4.2).

Table 4.2: Predicted probability of treatment given covariates by estimand

	Probability (min/max)		Weights
	Treated	Untreated	(min/max)
Estimand I	0.095 / 0.870	0.082 / 0.840	1.09 / 10.5
Estimand II	0.138 / 0.744	0.128 / 0.682	1.15 / 7.25
Estimand III (t=0)	0.148 / 0.790	0.083 / 0.730	1.09 / 6.74
Estimand III (t=1)	0.122 / 0.754	0.113 / 0.714	1.13 / 8.21

4.3.3 Estimation Results

Estimates for the point treatment effects and their corresponding confidence intervals are shown in table 4.3. The point estimates represent a difference in mean weight-for-age z-score. Therefore, a unit change of one is equivalent to one standard deviation above the mean weight for the reference standard (i.e., a population of well- nourished and healthy children of the same age and gender). In this section, I report the non-parametric confidence intervals (CI) based on the 2.5th and 97.5th percentiles of the bootstrap point estimates. The parametric CI from the bootstrap estimates (based on a normal distribution) and the influence curve-based CI from the original sample are shown in table 4.3. The distributions of the bootstrap estimates by estimand and estimation method are shown in the Appendix (A3). In all cases, the point estimate from the full sample falls within the range of the bootstrap estimates.

Estimand I: Post Treatment Outcome Y^c(t=1)

For estimand I, the parametric regression estimate of the average treatment effect (ATE) is small and not statistically significant at the 5% level ($\beta = 0.046$, CI: -0.031, 0.111). The effect estimate is reduced further with the use of IPTW, and the variance is comparable ($\beta = 0.038$, CI: -0.045, 0.086). The largest effect estimate is obtained with TMLE ($\beta = 0.066$, CI: 0.001, 0.146), which is consistent with TMLE adjusting more completely for negative confounding. Although the non-parametric CI is wider from TMLE than IPTW, the effect estimate from TMLE is statistically significant. The narrowest CI for Ψ^I, regardless of estimation method, is obtained from the TMLE influence curve (CI width = 0.116). Note that the differences in CI mentioned here and below, are very small.

Removing W^c(t=1) from the estimation reduces the point effect estimate by 13-20% for Ψ^I (i.e., from 0.066 down to 0.053 with TMLE) (see table 4.4). The width of the influence curve-based CI's for the estimates without W^c(t=1) remain essentially unchanged.

Estimand II: Change Score Outcome Y^0

The point estimate for estimand II from traditional regression is much larger than estimand I and statistically significant ($\beta = 0.276$, CI: 0.163, 0.376). The point estimates obtained from IPTW and TMLE are somewhat increased, although nearly identical, to those obtained from parametric regression. The estimates for Ψ^{II} remain statistically significant regardless of estimation method. The narrowest CI's for Ψ^{II} are the IPTW non-parametric CI and the TMLE influence curve (CI width = 0.181 and 0.180, respectively).

Table 4.3: Point treatment effect estimates and confidence intervals (includes $W^c(t=1)$)

Outcome	Target Parameter	Estimation Procedure	Point Estimate	Confidence Intervals (LCI, UCI)	Confidence Interval Method
Y(t=1)	Ψ^I	Parametric	0.046	-0.026, 0.118 -0.031, 0.111 -0.022, 0.114	Normal distr. Percentiles Influence curve
		IPTW	0.038	-0.028, 0.103 -0.045, 0.086 -0.040, 0.115	Normal distr. Percentiles Influence curve
		TMLE	0.066	-0.008, 0.142 0.001, 0.146 0.008, 0.124	Normal distr. Percentiles Influence curve
Y^θ	Ψ^{II}	Parametric	0.276	0.163, 0.390 0.163, 0.376 0.177, 0.375	Normal distr. Percentiles Influence curve
		IPTW	0.279	0.180, 0.378 0.167, 0.347 0.177, 0.381	Normal distr. Percentiles Influence curve
		TMLE	0.278	0.170, 0.387 0.137, 0.365 0.188, 0.369	Normal distr. Percentiles Influence curve
Y(t)	Ψ^{III}	Parametric	0.244	0.138, 0.351 0.146, 0.351 0.142, 0.346	Normal distr. Percentiles Influence curve
		IPTW	0.271	0.172, 0.369 0.166, 0.361 0.166, 0.375	Normal distr. Percentiles Influence curve
		TMLE	0.282	0.148, 0.416 0.196, 0.458 0.186, 0.378	Normal distr. Percentiles Influence curve

Table 4.4: Point treatment effect estimates and confidence intervals (excludes $W^c(t=1)$)

Outcome	Target Parameter	Estimation Procedure	Point Estimate	LCI, UCI †
Y(t=1)	Ψ^I	Parametric	0.037	-0.031, 0.104
		IPTW	0.033	-0.040, 0.107
		TMLE	0.053	-0.006, 0.114
Y^θ	Ψ^{II}	Parametric	0.249	0.149, 0.349
		IPTW	0.276	0.169, 0.384
		TMLE	0.277	0.166, 0.388

†Influence curve-based lower and upper confidence intervals (LCI, UCI)

Removing $W^c(t=1)$ from the estimation reduces the point effect estimates by 1-10% for Ψ^{II} (i.e., from 0.276 down to 0.249 with parametric regression). The width of the influence curve-based CI for TMLE increased from 0.181 to 0.222 (with no change in the point estimate).

Estimand III: Pooled Outcome $Y^c(t)$

As with estimand II, the point estimate for estimand III from traditional regression is relatively large and statistically significant ($\beta = 0.244$, CI: 0.146, 0.351). The effect estimate increases with the use of IPTW ($\beta = 0.271$, CI: 0.166, 0.361), and again with the use of TMLE ($\beta = 0.282$, CI: 0.196, 0.458). The estimates for Ψ^{III} remain statistically significant for all estimation methods. The narrowest CI for Ψ^{III} is obtained from the TMLE influence curve (CI width = 0.192). Note that it is not possible to test removing $W^c(t=1)$ from the estimation of Ψ^{III}.

4.4 Discussion

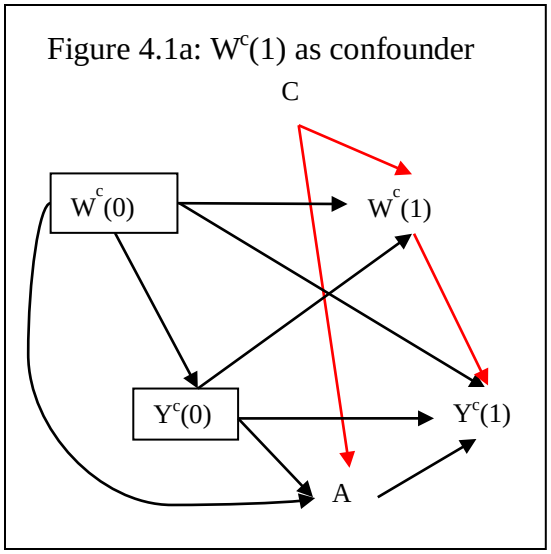
4.4.1 Estimation Results

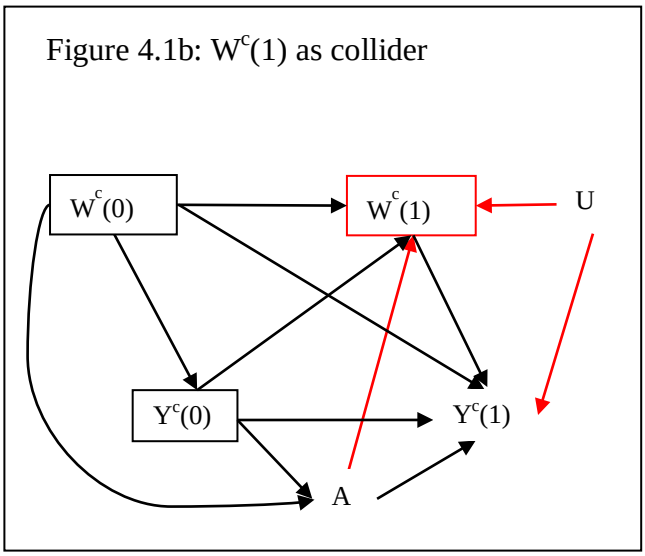
Estimand I: Post Treatment Outcome $Y^c(t=1)$

For estimand I to be equal to the ATE, it is sufficient that there are no unmeasured confounders, C (as discussed in chapter 3). Although this assumption is un-testable, it is possible that some unmeasured factors (such as community dispersion) are biasing the point estimate in either direction (i.e., the point estimate may be too small or too large). In order to accept that Ψ^I is equal to my parameter of interest, I must assume that such unmeasured confounders are weak and/or blocked by factors that I condition on.

All of the point estimates for Ψ^I are quite small: less than one tenth of a standard deviation in mean weight-for-age z-score. Only the TMLE estimate is statistically significant. The IPTW estimate is smaller than those obtained from traditional regression and TMLE, suggesting that IPTW may be biased downwards due to some practical violations of the ETA assumption. The traditional estimate is less sensitive to positivity violations, and is unbiased assuming no unmeasured confounding and correct parametric model specification for Ψ^I. However, given that this model is almost certainly mis-specified, the larger TMLE estimate for Ψ^I is the most reliable of the three estimators (does not rely on parametric models and is less sensitive to positivity violations).

The exclusion of $W^c(t=1)$ in the estimation did not increase the variance of any of the estimators (and inclusion of $W^c(t=1)$ does not improve the efficiency of the estimators). However, the 13 to 20% drop in the point estimates is worth considering in more depth. It is possible that there is an unmeasured factor that affects both A and $W^c(t=1)$ and opens a backdoor pathway to $Y^c(t=1)$ when I do not condition on $W^c(t=1)$ (i.e., a confounder, C, as shown in figure 4.1a). If this is the case, then it is best to condition on $W^c(t=1)$ to block this pathway. Alternatively, if A affects $W^c(t=1)$, then by conditioning on $W^c(t=1)$, I block the indirect effect of A on $Y^c(t=1)$ through $W^c(t=1)$, and I may be conditioning on a collider that opens up a pathway through some unmeasured common cause, U, of $W^c(t=1)$ and $Y^c(t=1)$ (see figure 4.1b). In this alternative scenario of A affects $W^c(t=1)$, it is best to not condition on $W^c(t=1)$. However, this alternative is less likely as it is contrary to background knowledge and the simple checks previously discussed.

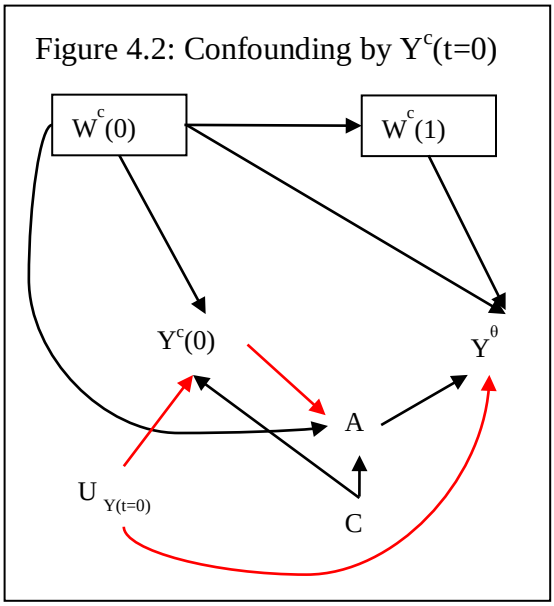




Estimand II: Change Score Outcome Y^{θ}

The change score estimand (Ψ^{II}) is potentially advantageous over the post treatment estimand (Ψ^{I}) in that it differences out unmeasured confounders, C, with a constant additive effect on $Y^c(t)$ at t=0,1. For example, if unmeasured community dispersion were a confounder, then its effect is eliminated in estimand II, if community dispersion is constant over time, and has a constant additive effect on WAZ. This cancelling effect is a compelling reason to consider estimand II and may explain the much larger effect estimate obtained for Ψ^{II}.

However, I demonstrate in chapter 3 that estimand II is equivalent to the ATE only under the RA that does not condition on $Y^c(t=0)$ and with the additional exclusion restrictions that $Y^c(t=0)$ does not affect A, $Y^c(t=1)$ and $W^c(t=1)$. As discussed previously, it is implausible that $Y^c(t=0)$ does not affect A. However, it is possible that $Y^c(t=0)$ does not affect $Y^c(t=1)$ and $W^c(t=1)$ except through $W^c(t=0)$ or V, given that the surveys were administered 7 years apart and were cross-sectional (i.e., different households and children were sampled in each period). Figure 4.2 represents a SCM where these two exclusion restrictions hold, but $Y^c(t=0)$ does, in fact, affect A. Under this model, confounding by $Y^c(t=0)$ occurs through exogenous $U_{Y(t=0)}$ (i.e., $A \leftarrow Y^c(t=0) \leftarrow U_{Y(t=0)} \rightarrow Y^{\theta}$). Therefore, in order to accept that Ψ^{II} is equal to my parameter of interest, I must assume that the residual variation in $Y^c(t=0)$ not explained by $W^c(t=0)$ and V has only a minimal influence on Y^{θ}.



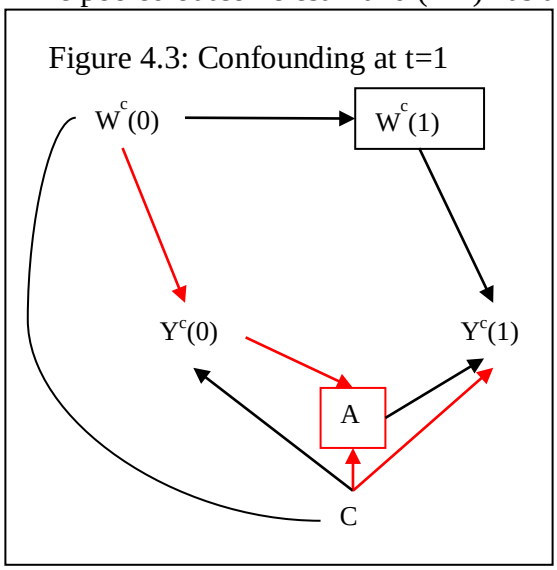
The point estimates for Ψ^{II} with all 3 estimators are nearly identical. In addition, Ψ^{II} is much less sensitive to removing $W^c(t=1)$ from the estimation than Ψ^{I} (impact is negligible with TMLE). This lack of differences in effect estimates suggests that Ψ^{II} is insensitive to the advantages of

using one estimator over another, and less sensitive to conditioning on $W^c(t=1)$. However, the wider influence curve-based CI for TMLE without $W^c(t=1)$ suggests that the inclusion of $W^c(t=1)$ improves the efficiency of the TMLE estimate for estimand II, most likely due to better prediction of Y^θ given covariates ($\overline{Q_0}$) using SuperLearner. In summary, I cannot be certain if the larger effect estimate for Ψ^{II} (in comparison to Ψ^{I}) represents an estimate of the ATE that is less biased by some unmeasured confounder or an estimate that is more biased due to not conditioning on $Y^c(t=0)$ (or possibly due to one of the other exclusion restrictions not holding).

Estimand III: Pooled Outcome $Y^c(t)$

The pooled outcome estimand (Ψ^{III}) has the same advantage as estimand II in that it differences out unmeasured confounders, C, with a constant additive effect on $Y^c(t)$ at t=0,1. However, for estimand III to be equal to the ATE, I must be willing to accept that $W^c(t=0)$ does not affect A and does not affect $Y^c(t=1)$, in addition to accepting the exclusion restrictions discussed above for estimand II. The characteristics in $W^c(t=0)$ are not known to have influenced treatment assignment, such that $W^c(t=0)$ does not affect A may be a reasonable assumption. It may also be plausible that $W^c(t=0)$ does not affect $Y^c(t=1)$ if, once again, I apply the logic that two cross-sectional samples separated by 7 years are only associated through fixed village characteristics, V. However, I found that mean infant birth rank in 1997 is negatively associated with mean weight-for-age in 2004 after conditioning on V, so empirically, there is a dependency. In addition, if I add that $Y^c(t=0)$ affects A in the SCM for estimand III at time t=1, then figure 4.3 demonstrates that $Y^c(t=1)$ is no longer independent of $W^c(t=0)$ given V, A, and $W^c(t=1)$ (a necessary condition). There is an unblocked pathway through collider A (i.e., $W^c(t=0) - C \rightarrow Y^c(t=1)$). To accept that Ψ^{III} is equal to my parameter of interest, I must be willing to accept that this dependency is negligible.

Figure 4.3: Confounding at t=1



The point estimate from parametric regression for Ψ^{III} is smaller than that obtained for Ψ^{II} (0.244 vs. 0.276). However, the effect estimate of Ψ^{III} with TMLE is the largest of all the estimates (0.282). Once again, I cannot be certain if the differences between the estimands is due to a less biased estimate of the ATE or due to bias from the additional assumptions for estimand III not holding. The CI for TMLE is wider for Ψ^{III} than for Ψ^{II}, suggesting that there was no efficiency gain with the pairing of $W^c(t)$ and $Y^c(t)$ at time t in estimand III.

In a prior analysis of this same dataset, the authors reported a mean effect on weight-for-age of approximately 0.218 using a traditional parametric regression at the individual level (as opposed to the community level).[43] Since the data represent different children at the two time points, only a re-formulation of estimand III at the individual level can support this type of analysis. The difference in point estimates obtained between the prior analysis at the individual level and my community level parametric estimate is small ($\beta = 0.218$ vs. 0.244). This difference may be explained by differences in the sample of communities used in the analysis and exact covariates

included in the regressions. Alternatively, an individual level analysis may gain precision, if the pairing of W(t) and Y(t) from the same individual at time t improves the prediction of the outcome.

The authors of the prior analysis recognize that the traditional difference-in-differences approach may be biased due to targeting of the poorest communities for receipt of the intervention. To address this concern, the authors used trimming to improve the exchangeability of the treated and untreated groups and aimed for an estimate of the average treatment effect among the treated (ATT). Specifically, they present an IPTW estimate ($\beta = 0.149$) where communities with propensity scores below the 5th and above the 95th percentiles were dropped from the analytic sample.[43] I cannot compare this estimate directly with my IPTW estimate because it represents a different target parameter, causal model, estimand, and identifiability result than I present here.

4.4.2 Conclusions

In summary, I am faced with a serious bias trade-off when choosing a final estimand for the ATE of the Madagascar nutrition program. The post treatment estimand (Ψ^{I}) controls for confounding due to the lagged outcome, $Y^{c}(t=0)$, but not from possible unmeasured confounder C. The change score and pooled outcome estimands (Ψ^{II} and Ψ^{III}) do not control for confounding by $Y^{c}(t=0)$, but have the potential to adjust for some types of unmeasured confounding. The extent to which unmeasured confounding is controlled depends on the specifics of the unmeasured confounder C (i.e., how C affects $Y^{c}(t=0)$ and $Y^{c}(t=1)$, and whether the effect of C changes over time). In addition, both estimands II and III have the potential for introducing bias if the additional assumptions they require (beyond estimand I) are not met. I am unable to estimate either the magnitude or direction of possible confounding from C, or the magnitude or direction of bias from the failure of the exclusion restriction assumptions to hold. Therefore, I conclude that my best choice is the post treatment estimand because it adjusts optimally for the known measured confounders and is equal to the ATE under the fewest assumptions.

Once the estimand is chosen, the choice of estimator can still make a difference. For Ψ^{I}, the largest effect estimate with the smallest variance was obtained with TMLE ($\beta = 0.066$, CI: 0.001, 0.146). I report the TMLE estimate, given the desirable properties of the method. TMLE has advantages over parametric regression in that it does not rely on correct model specification of $\overline{Q_0}$, but instead makes use of a non-parametric data-adaptive approach (SuperLearner) for prediction. TMLE has advantages over IPTW in that it is a substitution estimator that extrapolates into areas that lack support for experimentation (i.e., urban communities with high prevalence of underweight). TMLE improves on both of the other two methods by implementing a bias reduction step to estimate the target parameter of interest. Finally, TMLE is doubly robust to misspecification of either $\overline{Q_0}$ or g_0 and is maximally efficient if both are correctly specified.

In conclusion, I use my knowledge of the data generating system for the observed data and a detailed process of assumption checking to select a simple post treatment estimand for the average treatment effect of the Madagascar program on community mean weight-for-age. I follow this with a comparison of estimation methods and select the method for estimation that

provides the least biased estimate of the ATE. Combined, these choices result in a small, but statistically significant, estimate of benefit that can be attributed to the Madagascar program.

4.4.3 Future Work

Although I settle on a best estimate of the ATE given the observed data, I cannot rule out the possibility that the estimate is biased due to an unmeasured confounder, C. The reality of observational studies of large-scale programs is that there are tradeoffs that can lead to bias. The decision becomes one of choosing between approaches that lead to the most tolerable bias. Making this decision requires a full understanding of the choices. In this section, I consider alternate target parameters and alternate estimation approaches that might help to resolve the current uncertainty in the evaluation of the Madagascar program. In table 4.5, I summarize several possibilities for future work. I provide a brief comment for each about whether the alternative is likely to influence the trade-off (or the "$Y^c(t=0)$ dilemma") described here and in chapter 3. I consider several target parameters including extending the current analysis to the 2011 weight-for-age (WAZ) data, and the average treatment effect among the treated (ATT).

Table 4.5: Future evaluations of Madagascar's SEECALINE program

Alternate target parameters	**Comment**
ATE for mean WAZ in '11	An extension of the current analysis, no change to assumptions
Delayed ATE in '04 on WAZ in '11	An extension of the current analysis, no change to assumptions
ATT for mean WAZ in '04	Useful for addressing ETA violations, but does not resolve the $Y^c(t=0)$ dilemma and is unlikely to change the exclusion restrictions of the DiD estimators
Restrict dataset (i.e., drop communities)	
ATE for mean WAZ in '04 using individual as unit of analysis	Requires new SCM that takes into account the hierarchical nature of the data. Based on previously published results, it is not likely to meaningfully change the results
ATE for mean height-for-age (HAZ) in '04	Mean height as outcome may reduce the effect of the $Y^c(t=0)$ dilemma as HAZ is less predictive of treatment
Alternate estimation methods	
Propensity score matching	Typically applied to an ATT and to reduce the effect of measured confounding, but does not resolve the $Y^c(t=0)$ dilemma and has unknown effect on unmeasured confounding
Exact matching on $Y^c(t=0)$	The definition of an exact match is problematic and may have the same problem as conditioning on $Y^c(t=0)$ in a difference-in-differences estimator

The average treatment effect among the treated (ATT) is a conventional target parameter in the field of econometrics and impact evaluations of observational studies.[59] The ATT is of interest in the public health field as well, because it is a measure of effect among those who are most likely to need or seek treatment. A common approach to estimating the ATT is to select untreated communities that are comparable to those that are treated in terms of their conditional probability of treatment based on measured confounders.[66, 91] Communities are either trimmed from the analytic sample based on some fixed probability cut-points, or may be dropped in a

propensity score-matched analysis if no matches are found (although this depends on the method for matching).

Importantly, redefining the target causal parameter (as the ATT or other) does not resolve the problem of what to do with the lagged outcome, $Y^c(t=0)$, when it affects treatment assignment. And it is unlikely to change the other exclusion restrictions necessary for the popular difference-in-differences estimator to be equal to the target parameter (although this would need to be verified with steps 1 to 4 of the road map). Therefore, in order to pursue a change score or pooled outcome estimator, I will need a method that can avoid confounding bias due to the exclusion of the lagged outcome, $Y^c(t=0)$, from the conditioning subset. One possible method of interest is to match communities exactly on $Y^c(t=0)$.[60] However, the definition of an exact match may be problematic and prone to error from the village estimate. In addition, I may have the same problem with conditioning on $Y^c(t=0)$ as I do with a difference-in-differences estimator.

In summary, it is not clear that a methodological approach will resolve the bias trade-off. A final alternative may be to accumulate a consistent set of results that tell the same story. For example, a small but significant benefit to height-for-age in 2004 and a delayed benefit of treatment in 2004 on community nutritional outcomes in 2011 would provide supporting evidence of a causal benefit of the SEECALINE program.

Chapter 5: Conclusions

5.1 Overview

This dissertation focuses on several methodological issues in evaluating large-scale early child development (ECD) interventions that threaten the validity of estimating a program benefit. In chapter 2, I discuss the challenge of obtaining an unbiased measure of cognitive development in an ethnically diverse, low-income country setting. Using data from a study in Madagascar, I apply item response theory (IRT) models to assess the performance of a test of vocabulary knowledge, the Peabody Picture Vocabulary Test (PPVT). The IRT analysis uncovers problem areas and patterns of responses that can be used to identify items that need to be dropped before estimating the program effect (e.g., items with strong, significant DIF by local dialect), and to identify items that should be replaced, modified, or re-ordered in future work. I present lessons learned from working with the PPVT in Madagascar and make recommendations for how these lessons can be applied in other developing country settings. Specifically, when changing or dropping items are not allowed by the publisher, I recommend the following: a) pre-test many more items than are thought to be necessary; b) identify a minimum set of consecutive items with the best person separation reliability for a given age group; and c) administer this set of items without the use of start or stopping rules.

In chapters 3 and 4, I address the analytic challenges of determining whether an ECD intervention has a benefit that actually is the result of the intervention. I work my way through a roadmap for estimating the average treatment effect of Madagascar's national nutrition program on children's weight-for-age. I deliberately keep the process of definition and identification separate from the process of estimation in order to avoid confusion of the two. Throughout both chapters, I make the assumptions explicit and demonstrate the consequences of alternate choices, from the choice of the definition of the outcome to the method of estimation. These choices result in very different estimates of effect in the Madagascar study, where selection into the program was non-random and strongly associated with the pre-program outcome. The presentation style of the chapters is didactic, in the hopes that other investigators will be encouraged to follow the same basic steps when making analytic choices for their evaluations.

In this concluding chapter I reflect on some additional steps we (i.e., the community of ECD investigators) might take to improve our evaluations going forward. For example, in section 5.2, I speculate about the possibility of an open source model for sharing of information, test items and instruments for assessing cognition and language. In section 5.3, I urge investigators from separate disciplines to learn from, and incorporate, other disciplinary approaches to the identification of a causal parameter. And finally, in section 5.4, I encourage analysts to make use of the available estimation tools that meet the needs of their analytic goals.

5.2 Measurement – the case of language

Just in the last decade, the WHO reshaped how we think about children's growth potential around the world. The 2006 standards for height and weight set the bar for "how all children should grow rather than merely describing how children grew at a specified time and place."[108] Instead of being limited to comparisons within a population, we now compare growth across populations and advocate for change. Obtaining a similar set of international standards for cognition and language could like-wise be a powerful tool for change. Given my results on the performance of the PPVT in Madagascar, the reader might conclude that I think such a goal is

impossible to achieve. On the contrary, although it would be incredibly challenging, I think it is a necessary goal. Research in early child development is occurring across languages, cultures and countries (e.g., the Young Lives Study),[76] and the need already exists for making valid international comparisons.[109] I think it is possible with the right combination of tools, organization, and collective will.

Within the context of ECD research in developing countries today, I would categorize investigators into one of three (overly simplified) groups: 1) those who develop and validate their own tests; 2) those who translate/adapt an existing measure and validate it in some way; and 3) those who translate/adapt an existing measure and do not validate it. The first approach has important advantages in that the tests are culturally and linguistically relevant to the population being tested. However, developing new tests is resource intensive, both in terms of time and money, and will only occur where these resources are available. In addition, if the goal is to develop international standards, then country-specific tests will not get us there. For example, a group in Kilifi, Kenya[110] and another group in Hong Kong[111] have developed their own receptive vocabulary tests. Conceptually, these vocabulary tests are similar, but not identical to, the Peabody Picture Vocabulary Test (PPVT). Instead of using one target image with three distractors from the same category of word (e.g., running and three other play activities), the Kenyan and Cantonese instruments use one target picture, a phonological distractor, a visual or semantic distractor, and an unrelated distractor.[111] The inclusion of a phonological distractor increases the difficulty of transferring the test to another country: it is unlikely that one of the distractor images will sound the same as the stimulus word after translation.

The second approach of adapting and internally validating an existing measure may offer a way forward. Validation information for a given test gathered from different countries and languages could be made available to other interested parties (e.g., the Young Lives study has some publically available information).[76, 77] This sharing of information could be managed either through the publishers (e.g., Pearson for the PPVT) or through an open source model. The ultimate success of optimizing an instrument for a given context will depend on being able to modify, drop, add, and re-arrange the items as necessary. However, this may not be acceptable to the publisher of the test (depending on the test and the publisher's legal requirements). Currently, an open source model exists for early math and literacy assessment tools. Specifically, USAID has an organization, Education for Decision Making (EdData II), which has developed instruments (e.g., the EGMA and EGRA) that have been applied in 44 countries and in 80 languages.[112] Interested parties can download complete instruments and manuals in several languages, along with guidance for adaptations, final reports and presentations from studies conducted around the world. An alternate open source model that I envision would be one where for any given measure (e.g., vocabulary knowledge or matrix reasoning) an extensive library of items would be made available. In this model, investigators would choose culturally and age-appropriate items for their study, with the inclusion of a subset of items that would anchor the test for international comparisons and comparisons across age groups. Recommended methods of selecting, translating, testing and ordering items would need to be made available for first time use in a new language or country. Previously optimized collections of items would also be made available, as in the USAID model. Newly developed items, translations, and complete instrument assessments would be shared.

The results and discussion in chapter 2 make clear the scientific risk to inference from the third approach. Unfortunately, due to the lack of local gold standards for validation and/or the lack of familiarity with methods to internally validate the measures, the third approach is not uncommon. Although my work focuses on a test of vocabulary, many of the issues apply to any multi-item instrument intended to capture a latent construct. Such multi-item measures are commonly used in intervention research and include other tests of language and memory, as well as non-verbal tests of cognition and socio-emotional behavior scales. Two regional efforts (one in Latin America and one in East Asia-Pacific) are well underway to establishing comparative measures of children's early development across a broad cross-section of domains.[113, 114] Specifically, the Programa Regional de Indicadores de Desarrollo Infantil (PRIDI) is currently being tested in five Latin American countries[115] and the East Asia-Pacific Early Child Development (EAP-ECD) scales were scheduled for piloting in six Asian countries in 2011.[116] The stated tasks of the EAP-ECD research team include: "(i) creating a database of items and indicators by domain and age level across countries; (ii) analyzing and selecting a sub-group of items that are consistent across countries and represent a variety of domains; (iii) creating a final recommended list of items and domains between the 3 and 5 year old age groups; and (iv) preparing guidelines for the validation process."[114] In addition, there are international indicators for factors that influence early child development including UNICEF's Multiple Indicator Cluster Surveys (MICS),[117] which focuses on caregiving in the home, and FANTA's indicators for assessing infant and young child feeding practices.[118] Finally, UNESCO has a mandate to develop the holistic early childhood development index (HECDI), which will consider existing measures that assess factors influencing child development (such as the FANTA indicators and MICS), as well as indicators of achieving early developmental milestones (such as the PRIDI).[119] Acquiring a set of international standards for cognition and language may be an ambitious goal, but perhaps a necessary one given the needs of international research programs.[113-116, 119]

5.3 Identification of the target parameter – the case of pre/post data

Chapter 3 of my dissertation might be more accurately titled: "To ignore or not to ignore the unobservables? That is the question." Although different disciplines control for observed confounders in different ways, we all agree that we should control for them in the best way possible (the topic of chapter 4). However, disagreement runs deep with respect to those factors that we do not observe. In the Madagascar evaluation, I am faced with a bias trade-off between a single post-treatment estimand that conditions on the pre-treatment outcome (a measured confounder) but assumes no unmeasured confounders, and two difference-in-differences estimands that address certain types of unmeasured confounders but do not condition on the pre-treatment outcome. The following is a re-phrasing of text from an article that compares two methods (one from epidemiology, one from econometrics) in such a way that it speaks to this bias trade-off.[120] An economist will view the single post-treatment estimand with suspicion because the assumption of no unmeasured confounders seems unrealistic. On the other hand, an epidemiologist or biostatistician will be wary of the difference-in-differences estimands that purport to "subtract out" a variable that has not been (and possibly cannot be) observed. In reality, both rely on assumptions that cannot be empirically verified.

Given that the unobservables may also be unknown, how should I decide whether to ignore them or not? The authors of this same article have a take on the disciplinary differences that I find helpful:

> "… Many epidemiologic studies differ from those in social sciences in that the collection of candidate confounders is an integral part of study design. By contrast, important research questions in economics and the social sciences are usually addressed by analysing data that have been collected or are maintained by government agencies or survey organizations (e.g., Current Population Survey, Medicaid data, etc.). The databases serve as important resources for investigating a wide variety of issues, but the variables are not typically selected for a specific research agenda. Consequently, econometric methods for causal inference are predicated on the existence of at least one and possibly several unmeasured confounding variables; therefore, confounding is essentially viewed as an omitted variables problem that leads to correlation between errors and covariates (endogeneity)..."[120]

In the absence of a third alternative (to ignore, not to ignore, other?), I conclude that the answer to the question lies in part with the source of the data, and in part with what is known about the data generating system. I am convinced that the answer is inter-disciplinary and not specific to a given discipline. Consider the situation where information is collected on known confounders, for a well-defined research question, based on expert knowledge and the use of directed acyclic graphs (DAGs) or structural causal models (SCMs, which encode the same information). In this case, ignoring unobserved variables may be the best choice because the measurement of confounders was integral to the study and we can do a good job of controlling for them. We also avoid imposing additional untestable restrictions on the data. In the alternate situation where the research question is defined after the data were collected (e.g., from a national government survey), using a model that averages away the unobservables may be the better choice (again based on the research question, expert knowledge and DAGs).

Epidemiologists who use survey data collected for another purpose than their own can learn an important lesson from economists in choosing a statistical model for evaluation. For example, suppose an analyst wants to evaluate the effect of the national school lunch program on obesity among children in the U.S. using publically available data. The methods used by economists to evaluate labor or other policies may be necessary to account for unmeasured factors that may have influenced selection into the school lunch program. Similarly, economists who plan detailed measurement into their survey design *a priori* can learn an important lesson from epidemiologists. By incorporating DAGs into the planning process, researchers can identify a sufficient set of observables that should be measured to control for confounding. In this way, models that only hold under assumptions that may be implausible might be avoided.

5.4 Effect estimation – a comparison of methods

In chapter 4, I present three methods for estimation, and conclude that targeted maximum likelihood (TMLE) is a better choice over traditional parametric regression and inverse probability of treatment weighting (IPTW). Should the reader conclude that TMLE is the best choice for every research question? No; the answer depends on the question and what the investigator is trying to accomplish with their analysis. In my research, the answer is predicated on my goal of estimating a causal benefit of a nutrition intervention. In this section, I would like to discuss what the best choice of method might look like in other research contexts.

Most readers of this dissertation will know how to implement traditional parametric regression. In introductory statistics classes, we are taught how to interpret the coefficients on regression

terms. We know how to test for statistically significant associations between the covariates and the outcome, and we compare the relative strength and direction of these associations. The various relationships may help us to gain insight into the data generating system, to identify new confounders or effect modifiers or mediators, or to generate hypotheses for future research. If these are our stated goals, then estimation with parametric regression may be a good choice. Typically in this scenario, however, we report that the results are interpretable only as association, not causation.

In a second scenario, the coefficients from a parametric regression equation may be used for prediction. For example, I might want to create an algorithm to predict the probability of iron deficiency anemia given a child's age, gender, diet, etc. In this second scenario, I do not need to interpret any of the coefficients in a regression equation, instead I want to plug in values for each of the covariates (e.g., age = 5 months and breastfed = no) and get the best possible prediction of the outcome. Although it is possible to build a parametric algorithm for prediction, it is not necessarily the best choice. First, assuming that I have only one available sample, the data need to be split, with one part used for building the algorithm and the other for validating it. In other words, I need to test if the equation I just built actually does a good job of prediction, and I cannot use the same data to both build and validate it (referred to as over-fitting). Ideally, multiple splits and cross-validations would be performed, or multiple, separate samples would be used. Second, my goal is prediction and not interpretation of coefficients, so I can incorporate the predictor variables into my equation in any number of ways (e.g., as quadratic or interaction terms). I am also not limited to a linear regression algorithm: perhaps a non-parametric decision tree model would perform better for predicting my outcome. Very quickly, the process of building the best algorithm for prediction can become so onerous, that it is best to turn it over to a machine. This is exactly what the SuperLearner application is designed to do (see the appendix in chapter 4 for more detail), and in a prediction scenario, would be a much better choice than parametric regression.[102] This is also one of the reasons I use SuperLearner in the prediction steps for IPTW (to predict treatment given the covariates) and in TMLE (which has two prediction steps). The other reason is that the data-adaptive approach is non-parametric and avoids imposing unnecessary parametric assumptions on my model of the program effect.

In a third scenario, a single coefficient (or maybe 2) in a parametric regression may, in fact, be the solution to my research question. The rest of the covariates are parameters that are included in the regression only to control for confounding. For example, in a regression on the incidence of malaria (i.e., the outcome), the coefficient on the use of bed nets (i.e., treatment) may be the estimate I want to report, with age, gender, household size, socio-economic status and geographic location included as confounders. This one coefficient on treatment in a parametric equation may be equal to my causal parameter of interest (if I'm lucky), but there are a number of reasons why it may not be, which I discuss in detail in chapter 4. Importantly, parametric regression imposes assumptions about the functional form of the covariates and their relationship with the exposure and outcome (e.g., linear and additive) that may not hold. If the equation is incorrectly specified, then my causal effect estimate will be biased. In addition, parametric regression is designed to obtain the best fit for the full conditional distribution of the outcome given the exposure and covariates (as is SuperLearner). However, I am not interested in this full distribution (or all the coefficients on the covariate terms). Instead, I want the best possible estimate of the mean difference in the outcome under treatment and no treatment (for the ATE).

Since parametric regression does not give me the optimal fit for my actual parameter of interest, the estimate will be biased.[58] TMLE resolves both of these shortcomings by using SuperLearner for two prediction steps and implementing a bias reduction step that targets my parameter of interest. TMLE has additional advantages (such as double robustness) that make it a better choice over parametric regression and IPTW for my research.

In conclusion, I would like to tie this choice of estimation method discussion back to the question of ignorability. Any of the three estimation methods I present (and more) can be used with any number of statistical models, regardless of whether the answer is to ignore or not to ignore the unobservables. The choice of estimator, such as a difference-in-differences estimator, should not be confused with the choice of method of estimation. Both choices are important for obtaining an unbiased causal effect estimate, which is an ambitious goal with which to begin. I strongly recommend that investigators work through the road map (or something comparable) and use DAGs or SCMs (or both) before choosing the estimator. I also recommend implementing TMLE for the method of estimation of causal effects, despite its seeming complexity. Development of applications for TMLE is on-going, and the method has been successfully applied to a broad cross-section of health related research.[58] Both SuperLearner and TMLE packages are available to download for free from CRAN (the comprehensive R archive network) and TMLE can be run with little effort using the package default settings (see R Package *tmleLite* version 1.0-2). The evaluation of a causal benefit of an intervention (or a policy) on an outcome from observational data is a bold goal, but one for which an increasing number of methods are available to help us achieve.

5.5 Final Remarks

In my dissertation, I present several challenges associated with estimating a causal benefit of a large-scale ECD intervention. I present how these challenges can be tackled and urge investigators to update and/or reconsider their analytic approaches to evaluations. I recommend using the methods that are at our disposal, to learn new methods from other disciplines, and to not hold on too tightly to our own disciplinary methods. Otherwise, we run the risk of estimating a program effect that is misleading, which may negatively affect those whom the program is intended to benefit. Finally, in this last chapter, I suggest setting ambitious goals for moving forward. These include creating a set of international standards for testing and applying a common set of methods for causal inference.

References

1. Grantham-McGregor S, Cheung YB, Cueto S, Glewwe P, Richter L, Strupp B, et al. Developmental potential in the first 5 years for children in developing countries. Lancet. 2007;369(9555):60-70.
2. Walker SP, Wachs TD, Meeks Gardner J, Lozoff B, Wasserman GA, Pollitt E, et al. Child development: risk factors for adverse outcomes in developing countries. The Lancet. 2007 2007/1/19/;369(9556):145-57.
3. Walker SP, Wachs TD, Grantham-McGregor S, Black MM, Nelson CA, Huffman SL, et al. Inequality in early childhood: risk and protective factors for early child development. The Lancet. 2011;378(9799):1325-38.
4. Engle PL, Fernald LCH, Alderman H, Behrman J, O'Gara C, Yousafzai A, et al. Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries. The Lancet. 2011;378(9799):1339-53.
5. Victora CG, Adair L, Fall C, Hallal PC, Martorell R, Richter L, et al. Undernutrition 2: Maternal and Child Undernutrition: Consequences for Adult Health and Human Capital. The Lancet. 2008;371(9609):340-57.
6. Black RE, Allen LH, Bhutta ZA, Caulfield LE, de Onis M, Ezzati M, et al. Maternal and child undernutrition: global and regional exposures and health consequences. The Lancet. 2008 2008/1/25/;371(9608):243-60.
7. Imdad A, Sadiq K, Bhutta ZA. Evidence-based prevention of childhood malnutrition. Current Opinion in Clinical Nutrition & Metabolic Care. 2011;14(3):276-85 10.1097/MCO.0b013e328345364a.
8. Bryan J, Osendarp S, Hughes, Calvaresi E, Baghurst K, van Klinken J-W. Nutrients for Cognitive Development in School-aged Children. 2004;62:295-306.
9. Beard JL. Why Iron Deficiency Is Important in Infant Development. J Nutr. 2008 December 1, 2008;138(12):2534-6.
10. UNICEF. Info by Country: Madagascar Statistics. New York, NY2011 [cited 2011 February 9]; Available from: http://www.unicef.org/infobycountry/madagascar_statistics.html.
11. Sharp M, Kruse I. Health, nutrition, and population in Madagascar 2000-09. Africa Human Development Series. Washington D.C.: The World Bank; 2011.
12. Galasso E. Personal communication. 2011.
13. Shonkoff JP, Phillips DA, editors. From Neurons to Neighborhoods: The Science of Early Childhood. Development Committee on Integrating the Science of Early Childhood Development. Washington D.C.: National Academy Press; 2000.
14. Casey BJ, Tottenham N, Liston C, Durston S. Imaging the developing brain: what have we learned about cognitive development? Trends in Cognitive Sciences. 2005;9(3):104-10.
15. Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life Course Epidemiology. Journal of Epidemiology and Community Health (1979-). 2003;57(10):778-83.
16. Victora CG, de Onis M, Hallal PC, Blössner M, Shrimpton R. Worldwide Timing of Growth Faltering: Revisiting Implications for Interventions. Pediatrics. 2010 February 15, 2010.
17. Stein AD, Wang M, Martorell R, Norris SA, Adair LS, Bas I, et al. Growth patterns in early childhood and final attained stature: Data from five birth cohorts from low- and middle-income countries. American Journal of Human Biology. 2010;22(3):353-9.
18. Kuhn D, Siegler R. Handbook of Child Psychology. 5 ed. New York1998.

19. Bryan J, Osendarp S, Hughes D, Calvaresi E, Baghurst K, Klinken J-W. Nutrients for Cognitive Development in School-aged Children. Nutrition Reviews. 2004;62(8):295-306.
20. Beckett C, Maughan B, Rutter M, Castle J, Colvert E, Groothues C, et al. Do the Effects of Early Severe Deprivation on Cognition Persist Into Early Adolescence? Findings from the English and Romanian Adoptees Study. Child Development. 2006;77(3):696-711.
21. Hertzman C, Boyce T. How Experience Gets Under the Skin to Create Gradients in Developmental Health. Annual Review of Public Health. 2010;31(1):329-47.
22. Heckman JJ, Stixrud J, Urzua S. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Journal of Labor Economics. 2006;24(3):411-82.
23. Alderman H. Improving Nutrition through Community Growth Promotion: Longitudinal Study of the Nutrition and Early Child Development Program in Uganda. World Development. 2007;35(8):1376-89.
24. Popkin BM, Canahuati J, Bailey PE, O'Gara C. An evaluation of a national breast-feeding promotion programme in Honduras. Journal of Biosocial Science. 1991;23(01):5-21.
25. Bhutta ZA, Ahmed T, Black RE, Cousens S, Dewey K, Giugliani E, et al. What works? Interventions for maternal and child undernutrition and survival. The Lancet. 2008;371:417-40.
26. Engle PL, Black MM, Behrman JR, Cabral de Mello M, Gertler PJ, Kapiriri L, et al. Strategies to avoid the loss of developmental potential in more than 200 million children in the developing world. The Lancet. 2007;369(9557):229-42.
27. Paxson C, Schady N. Does Money Matter? The Effects of Cash Transfers on Child Health and Development in Rural Ecuador. Economic Development and Cultural Change. 2008;59(1):187-229.
28. Fernald LCH, Gertler PJ, Neufeld LM. Role of cash in conditional cash transfer programmes for child health, growth, and development: an analysis of Mexico's Oportunidades. The Lancet. 2008 2008/3/14/;371(9615):828-37.
29. Macours K, Schady N, Vakis R. Cash Transfers, Behavioral Changes, and the Cognitive Development of Young Children: Evidence from a Randomized Experiment. Washington, DC: World Bank2008.
30. Armecin G, Behrman JR, Duazo P, Ghuman S, Gultiano S, King EM, et al. Early childhood development through an integrated program : evidence from the Philippines. Policy Research Working Paper Series 2006(3922).
31. Galasso E, Umapathi N. Improving nutritional status through behavioral change : lessons from Madagascar World Bank Policy Research Working Paper Series No 4424. Washington DC2007.
32. Roberfroid D, Kolsteren P, Hoerée T, Maire B. Do growth monitoring and promotion programs answer the performance criteria of a screening program? A critical analysis based on a systematic review. Tropical Medicine & International Health. 2005;10(11):1121-33.
33. Ashworth A, Shrimpton R, Jamil K. Growth monitoring and promotion: review of evidence of impact. Maternal & Child Nutrition. 2008;4:86-117.
34. Huffman SL, Harika RK, Eilander A, Osendarp SJM. Essential fats: how do they affect growth and development of infants and young children in developing countries? A literature review. Maternal & Child Nutrition. 2011;7:44-65.
35. Michaelsen KF, Lauritzen L, Mortensen EL. Effects of Breast-feeding on Cognitive Function. In: Goldberg G, Prentice A, Prentice A, Filteau S, Simondon K, editors. Breast-Feeding: Early Influences on Later Health: Springer Netherlands; 2009. p. 199-215.

36. Kramer MS, Aboud F, Mironova E, Vanilovich I, Platt RW, Matush L, et al. Breastfeeding and Child Cognitive Development: New Evidence From a Large Randomized Trial. Arch Gen Psychiatry. 2008;65(5):578-84.
37. Kramer MS, Matush L, Vanilovich I, Platt RW, Bogdanovich N, Sevkovskaya Z, et al. Effects of prolonged and exclusive breastfeeding on child height, weight, adiposity, and blood pressure at age 6.5 y: evidence from a large randomized trial. The American Journal of Clinical Nutrition. 2007 December 1, 2007;86(6):1717-21.
38. Dewey KG, Adu-Afarwuah S. Systematic review of the efficacy and effectiveness of complementary feeding interventions in developing countries. Maternal & Child Nutrition. 2008;4(s1):24-85.
39. Lutter CK, Rodríguez A, Fuenmayor G, Avila L, Sempertegui F, Escobar J. Growth and Micronutrient Status in Children Receiving a Fortified Complementary Food. The Journal of Nutrition. 2008 February 1, 2008;138(2):379-88.
40. Hossain SMM, Duffield A, Taylor A. An evaluation of the impact of a US$60 million nutrition programme in Bangladesh. Health Policy and Planning. 2005 January 2005;20(1):35-40.
41. Gartner A, Kameli Y, Traissac P, Dhur A, Delpeuch F, Maire B. Has the first implementation phase of the Community Nutrition Project in urban Senegal had an impact? Nutrition (Burbank, Los Angeles County, Calif). 2007;23(3):219-28.
42. Dunn LM, Dunn LM. Peabody Picture Vocabulary Test - Third Edition. Circle Pines, MN: American Guidance Services, Inc; 1997.
43. Galasso E, Umapathi N. Improving nutritional status through behavioural change: lessons from Madagascar. Journal of Developmental Effectiveness. 2009;1(1):60-85.
44. Fernald LCH, Kariger PK, Engle P, Raikes A. Examining child development in low-income countries: A toolkit for the assessment of children in the first five years of life. Washington D.C.: The World Bank; 2009.
45. Fenson L, Dale PS, Reznick JS, Thal D, Bates E, Hartung JP, et al. The MacArthur Communicative Development Inventories: User's Guide and Technical Manual. Baltimore: Paul H. Brokes Publishing Co.; 1993.
46. Pollitt E, Triana N. Stability, predictive validity, and sensitivity of mental and motor development scales and pre-school cognitive tests among low-income children in developing countries. Food and Nutrition Bulletin. 1999;20(1).
47. Paxson C, Schady N. Cognitive development among young children in Ecuador : the roles of wealth, health, and parenting. The Journal of Human Resources. 2007;XLII(1):49-84.
48. Fernald LCH, Weber A, Galasso E, Ratsifandrihamanana L. Socioeconomic gradients and child development in a very low income population: evidence from Madagascar. Developmental Science. 2011;14(4):832-47.
49. American Psychological Association. The Standards for Educational and Psychological Testing. 1999 [29 November 2011]; Available from: www.apa.org/science/programs/testing/standards.aspx.
50. Williams KT, Wang J-J. Technical References to the Peabody Picture Vocabulary Test - Third Edition (PPBT-III). Circle Pines, MN: American Guidance Service, Inc.; 1997.
51. Marfo K, Pence A, LeVine RA, LeVine S. Strengthening Africa's Contributions to Child Development Research: Introduction. Child Development Perspectives. 2011;5(2):104-11.
52. Nores M, Barnett WS. Benefits of early childhood interventions across the world: (Under) Investing in the very young. Economics of Education Review. 2010 April;29(2):271-82.

53. Peña ED. Lost in Translation: Methodological Considerations in Cross-Cultural Research. Child Development. 2007;78(4):1255-64.
54. Sanchez A. Early nutrition and later cognitive achievement in developing countries. Paris: UNESCO; 2009.
55. Grantham-McGregor SM, Walker SP, Chang SM, Powell CA. Effects of early childhood supplementation with and without stimulation on later development in stunted Jamaican children. Am J Clin Nutr. 1997 Aug;66(2):247-53.
56. Wilson M, Allen DD, Li JC. Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach. Health Education Research. 2006 December 1, 2006;21(suppl 1):i19-i32.
57. Baranowski T, Allen DD, Masse LC, Wilson M. Does participation in an intervention affect responses on self-report questionnaires? Health Education Research. 2006 December 1, 2006;21(suppl 1):i98-i109.
58. van der Laan MJ, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. Berlin Heidelberg New York: Springer; 2011.
59. Heckman JJ, Vytlacil EJ. Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. In: James JH, Edward EL, editors. Handbook of Econometrics: Elsevier; 2007. p. 4779-874.
60. Imai K. Causal Inference Lecture Notes: Causal Inference with Repeated Measures in Observational Studies. Department of Politics, Princeton University2008.
61. Meyer BD. Natural and Quasi-Experiments in Economics. National Bureau of Economic Research Technical Working Paper Series. 1994;170.
62. Gertler PJ, Martinez S, Premand P, Rawlings LB, Vermeersch CMJ. Impact evaluation in practice. Washington DC: The International Bank for Reconstruction and Development / The World Bank; 2011.
63. Lord FM. The Measurement of Growth. Educational and Psychological Measurement. 1956 December 1, 1956;16(4):421-37.
64. Robins JM, Hernán MÁ, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000;11(5):550-60.
65. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. Journal of Epidemiology and Community Health. 2006;60:578-86.
66. Hirano K, Imbens GW, Ridder G. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. Econometrica. 2003;71(4):1161-89.
67. van der Laan MJ, Rubin D. Targeted Maximum Likelihood Learning. The International Journal of Biostatistics. 2006;2(1).
68. Mark J. van der Laan, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer; 2011.
69. Traub RE. Classical Test Theory in Historical Perspective. Educational Measurement: Issues and Practice. 1997;16(4):8-14.
70. Paek I, Wilson M. Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel–Haenszel Procedure in Short Test and Small Sample Conditions. Educational and Psychological Measurement. 2011 December 1, 2011;71(6):1023-46.
71. Wilson M, Allen DD, Li JC. Improving measurement in health education and health behavior research using item response modeling: Introducing item response modeling. Health Education Research. 2006 December 1, 2006;21(suppl 1):i4-i18.

72. Ferne T, Rupp AA. A Synthesis of 15 Years of Research on DIF in Language Testing: Methodological Advances, Challenges, and Recommendations. Language Assessment Quarterly. 2007 2007/07/03;4(2):113-48.
73. Restrepo MA, Schwanenflugel PJ, Blake J, Neuharth-Pritchett S, Cramer SE, Ruston HP. Performance on the PPVT-III and the EVT: Applicability of the Measures With African American and European American Preschool Children. Lang Speech Hear Serv Sch. 2006 January 1, 2006;37(1):17-27.
74. Wikipedia. Malagasy Language. [cited 2012 March]; Available from: http://en.wikipedia.org/wiki/Malagasy_language.
75. Encyclopedia of the Nations. Madagascar. [cited 2012 March]; Available from: http://www.nationsencyclopedia.com/Africa/Madagascar.html.
76. Cueto S, Leon J, Guerrero G, Muñoz I. Psychometric characteristics and cognitive development and achievement instruments in Round 2 of Young Lives. Young Lives Technical Note #15. . Oxford: Department of International Development; 2009.
77. Cueto S, Leon J, Guerrero G, Muñoz I. Annex 1: Item Statistics. Psychometric characteristics and cognitive development and achievement instruments in Round 2 of Young Lives Young Lives Technical Note #15. Oxford: Department of International Development; 2009.
78. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, CA: Sage Publications, Inc.; 1991.
79. Wilson M, Zheng X, McGuire L. Formulating Latent Growth using an Explanatory Item Response Model Approach. 2011.
80. Adams RJ, Wilson M, Wang W-C. The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement. 1997;21(1):1-23.
81. Spearman C. The Proof and Measurement of Association between Two Things. The American Journal of Psychology. 1904;15(1):72-101.
82. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334.
83. Wu ML, Adams RJ, Wilson MR, Haldane SA. ACER ConQuest version 2.0: Generalized Item Response Modelling Software. Victoria, Australia: ACER Press; 2007.
84. Wilson M. Constructing Measures: An Item Response Modeling Approach. Mahway, NJ: Lawrence Erlbaum Associates; 2005.
85. Paek I. Investigation of differential item functioning: Comparisons among approaches, and extension to a multidimensional context. : University of California at Berkeley; 2002.
86. Spearman C. Correlation calculated from faulty data. British Journal of Psychology, 1904-1920. 1910;3(3):271-95.
87. Muchinsky PM. The Correction for Attenuation. Educational and Psychological Measurement. 1996 February 1, 1996;56(1):63-75.
88. The World Bank. Ratio of girls to boys in primary and secondary education (%) Washington DC2009 [cited 2012 April]; Available from: http://data.worldbank.org/indicator/SE.ENR.PRSC.FM.ZS.
89. Paxson C, Schady N. Cognitive Development among Young Children in Ecuador. Journal of Human Resources. 2007 December 21, 2007;XLII(1):49-84.
90. Simos PG, Sideridis GD, Protopapas A, Mouzaki A. Psychometric Evaluation of a Receptive Vocabulary Test for Greek Elementary Students. Assessment for Effective Intervention. 2011 July 18, 2011.

91. Rubin DB. Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies. Journal of Educational Psychology. 1973;66:688-701.
92. Pearl J. Causal diagrams for empirical research. Biometrika. 1995 December 1, 1995;82(4):669-88.
93. Pearl J. An Introduction to Causal Inference. The International Journal of Biostatistics. 2010;6(2).
94. Maris E. Covariance adjustment versus gain scores—revisited. Psychological Methods. 1998;3(3):309-27.
95. Greenland S, Pearl J, Robins JM. Causal Diagrams for Epidemiologic Research. Epidemiology. 1999;10(1):37-48.
96. Petersen ML, Porter KE, Gruber S, Wang Y, Laan MJvd. Positivity. In: M.J. van der Laan, Rose S, editors. Targeted Learning: Causal Inference for Observational and Experimental Data. Berlin Heidelberg New York: Springer; 2011.
97. Rubin DB, Stuart EA, Zanutto EL. A Potential Outcomes View of Value-Added Assessment in Education. Journal of Educational and Behavioral Statistics. 2004;29(1):103-16.
98. Allison PD. Change scores as dependent variables in regression analysis. Sociological Methodology. Oxford & Cambridge, MA: Basil Blackwell Ltd.; 1990. p. 93-114.
99. Angrist J, Pischke J-S. Mostly harmless econometrics: an empiricist's companion. Princeton University Press; 2009.
100. Guryan J. Desegregation and Black Dropout Rates. The American Economic Review. 2004;94(4):919-43.
101. Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural Models. American Journal of Epidemiology. 2008 September 15, 2008;168(6):656-64.
102. Polley EC, Rose S, Laan MJvd. Super Learning. In: M.J. van der Laan, Rose S, editors. Targeted Learning: Causal Inference for Observational and Experimental Data. Berlin Heidelberg New York: Springer; 2011.
103. WHO Multicentre Growth Reference Study Group, de Onis M, Martorell R, Garza C, Lartey A. WHO Child Growth Standards based on length/height, weight and age. Acta Pædiatrica. 2006;95(Supp 450):76-85.
104. Wang Y, Petersen ML, Bangsberg D, Mark J. van der Laan. Diagnosing Bias in the Inverse Probability of Treatment Weighted Estimator Resulting from Violation of Experimental Treatment Assignment. UC Berkeley Division of Biostatistics Working Paper Series [serial on the Internet]. 2006: Available from: http://www.bepress.com/ucbbiostat/paper211
105. Glymour MM, Weuve J, Berkman LF, Kawachi I, Robins JM. When Is Baseline Adjustment Useful in Analyses of Change? An Example with Education and Cognitive Change. American Journal of Epidemiology. 2005;162(3):267-78.
106. Yanez III ND, Kronmal RA, Shemanski LR, Psaty BM. A Regression Model for Longitudinal Change in the Presence of Measurement Error. Annals of Epidemiology. 2002;12:34-8.
107. Chambless LE, Roeback JR. Methods for assessing difference between groups in change when initial measurement is subject to intra-individual variation. Statistics in Medicine. 1993;12:1213-37.
108. WHO Multicentre Growth Reference Study Group, Garza C, Onis Md, Martorell R, Onyango AW, Victora CG, et al. Assessment of differences in linear growth among populations in the WHO Multicentre Growth Reference Study. Acta Pædiatrica. 2006;95(Supp 450):56-65.

109. Fernald LCH, Kariger P, Hidrobo M, Gertler P. Socio-economic gradients in child development in very young children: Evidence from India, Indonesia, Peru and Senegal. 2012(Manuscript submitted for publication).
110. Holding PA, Taylor HG, Kazungu SD, Mkala T, Gona J, Mwamuye B, et al. Assessing cognitive outcomes in a rural African population: development of a neuropsychological battery in Kilifi district, Kenya. 2004.
111. Cheung PSP, Lee KYS, Lee LWT. The development of the 'Cantonese Receptive Vocabulary Test' for children aged 2–6 in Hong Kong. International Journal of Language & Communication Disorders. 1997;32(1):127-38.
112. USAID. EdData II: Education data for decision making. Research Triangle Park, North Carolina RTI International; [April, 2012]; Available from: https://www.eddataglobal.org/about/index.cfm.
113. Verdisco A. Without Data, There is No Action: Regional Project on Child Development Indicators, PRIDI. [25 April 2012]; Available from: http://www.iadb.org/en/topics/education/without-data-there-is-no-action,7454.html.
114. Asia-Pacific Regional Network for Early Childhood (ARNEC). East Asia-Pacific ECD Scale Development Workshop. ARNEC; 2011 [25 April 2012]; Available from: http://www.arnec.net/cos/o.x?ptid=1036089&c=/swt_arnec/articles&func=view&rid=253.
115. Fernald LH. Personal communication. 2012.
116. Asia-Pacific Regional Network for Early Childhood (ARNEC). Phase III: Next Steps for the East Asia-Pacific ECD Scales Development. ARNEC; 2011 [25 April 2012]; Available from: http://www.arnec.net/cos/o.x?ptid=1036089&c=/swt_arnec/articles&func=view&rid=260.
117. UNICEF. Multiple Indicator Cluster Survey (MICS). [25 April 2012]; Available from: http://www.unicef.org/statistics/index_24302.html.
118. Food and Nutrition Technical Assistance (FANTA). Indicators for assessing infant and young child feeding practices series (2008-2010). FANTA; [25 April 2012]; Available from: http://www.fantaproject.org/publications/iycf_definitions2008.shtml.
119. Tinajero AR, Loizillon A. Review of care, education and child development indicators in early childhood: United Nations Educational Scientific and Cultural Organization (UNESCO)2012.
120. Hogan JW, Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. Statistical Methods in Medical Research. 2004 February 1, 2004;13(1):17-48.

Appendices

A1: Additional Figures and Statistics for IRT Study

Figure 2.7: Raw score distributions at 2 time points

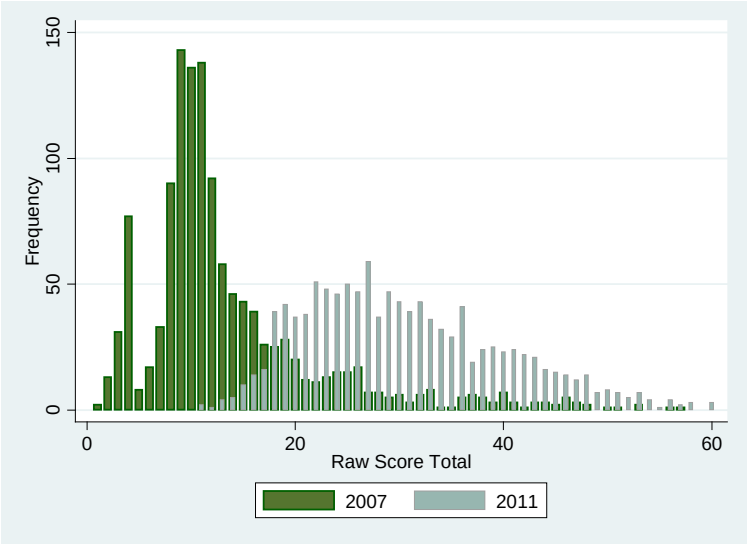


Figure 2.8: Separate, unidimensional IRT score distributions at 2 time points

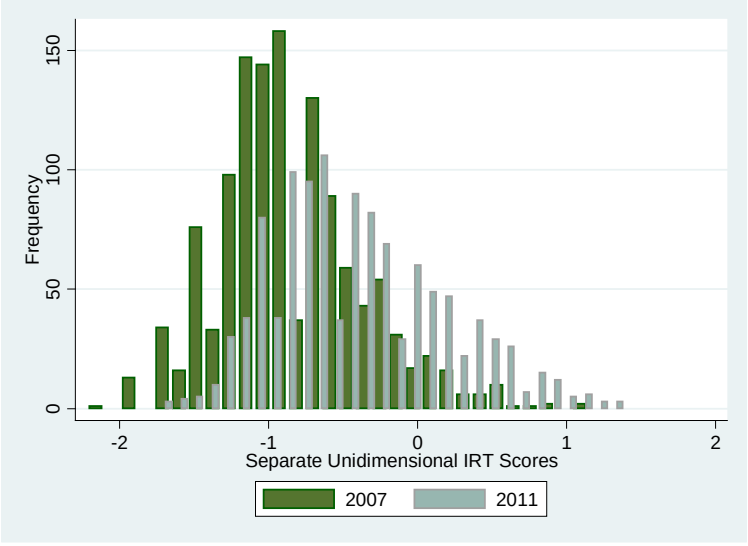


Figure 2.9: 2007 Wright Map – All children, all items, unidimensional IRT model

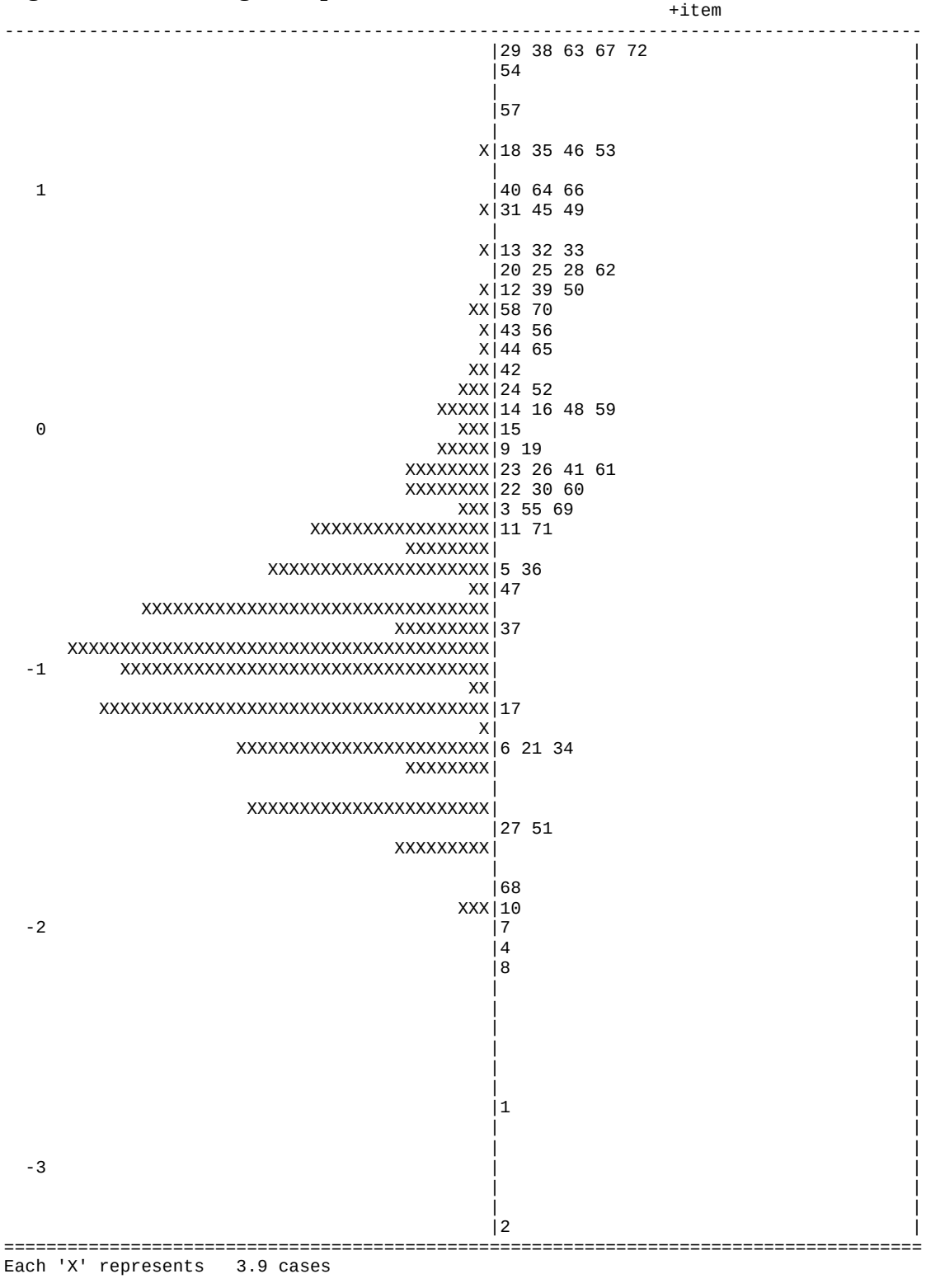


Figure 2.10: 2011 Wright Map – All children, all items, unidimensional IRT model

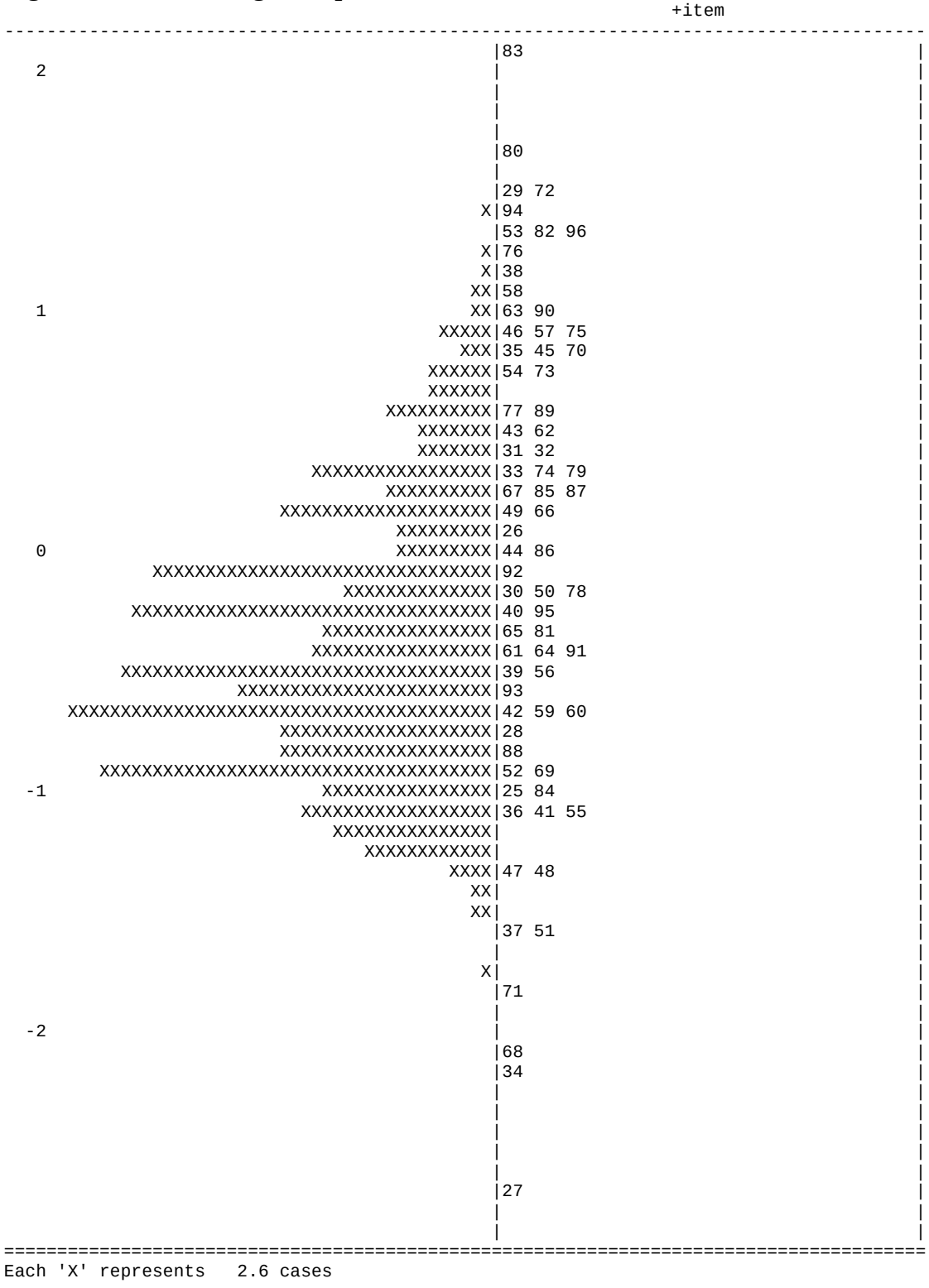


Figure 2.11: '07/'11 Wright Map – All children, all items, multidimensional IRT model

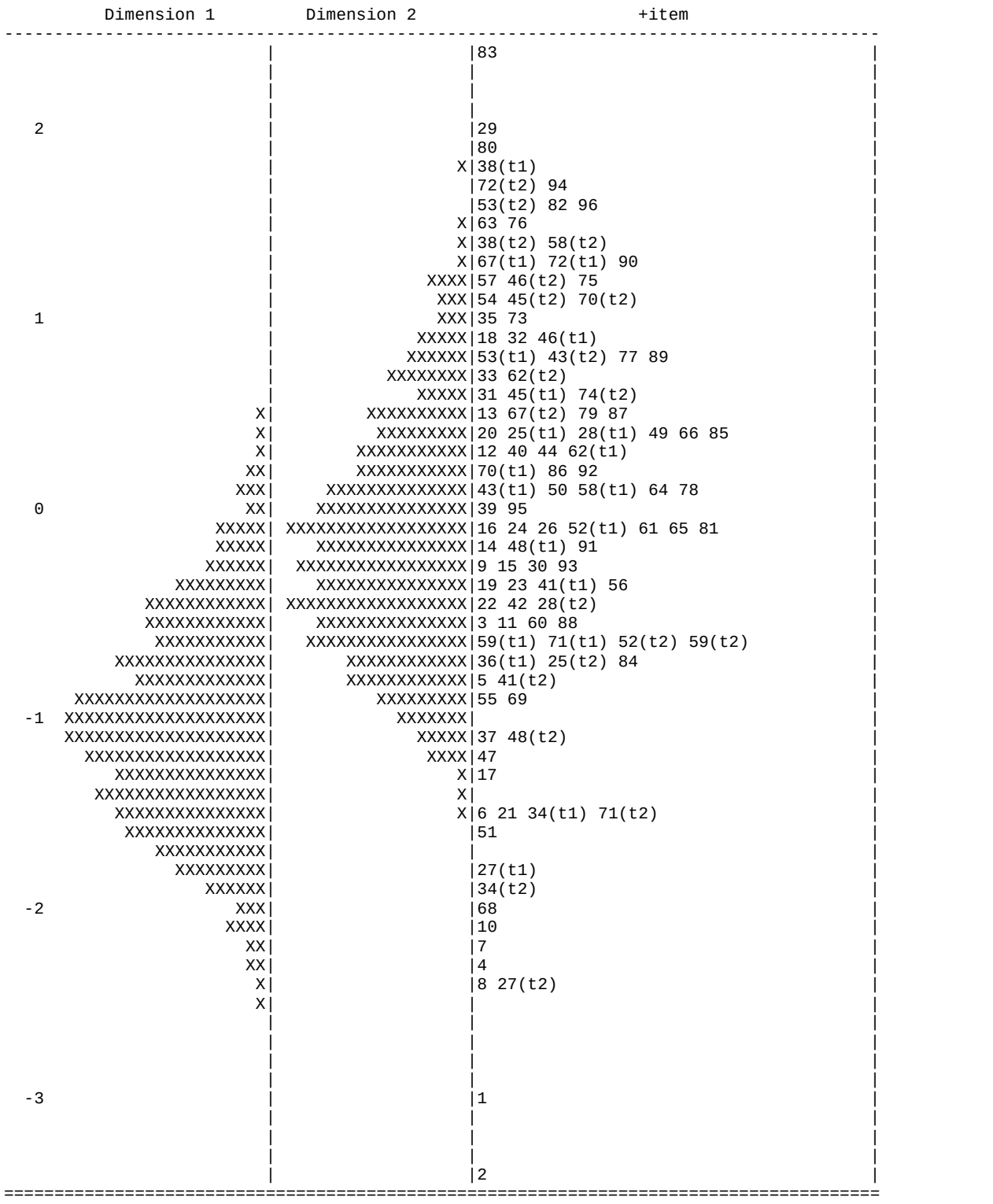


Figure 2.12: Item infit statistics by year - All children, all items, separate unidimensional IRT models

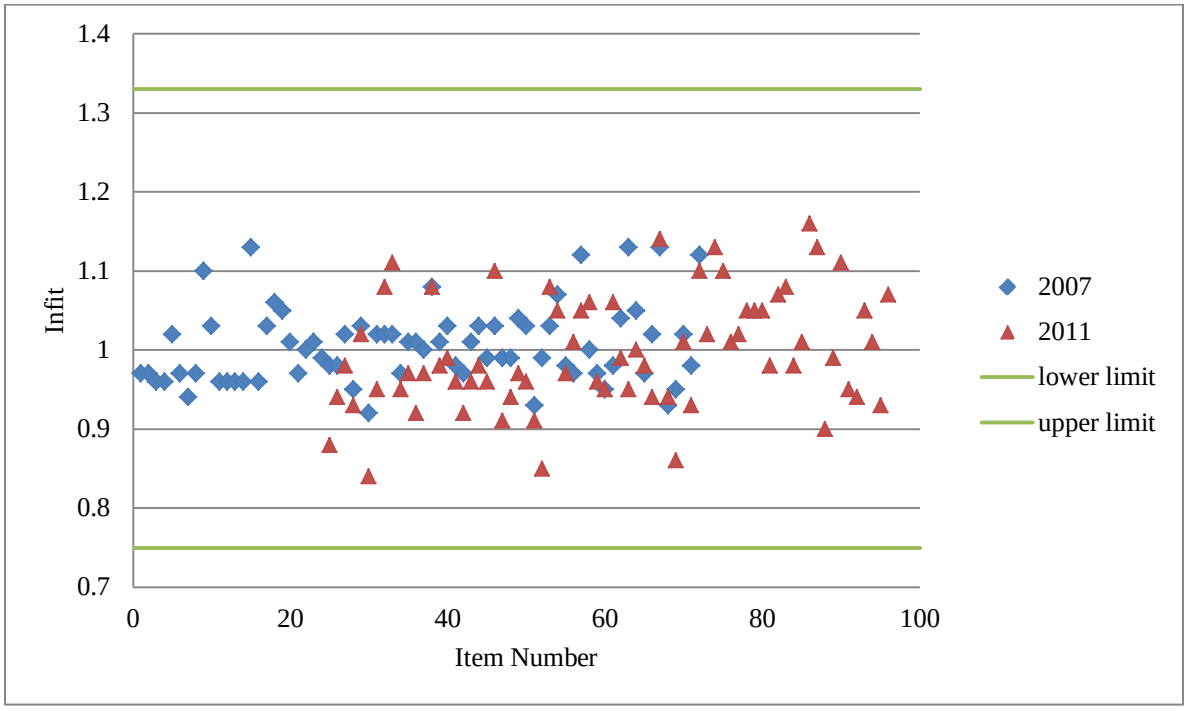


Figure 2.13: Item infit statistics by year - All children, all items, 2-dimenstional IRT model

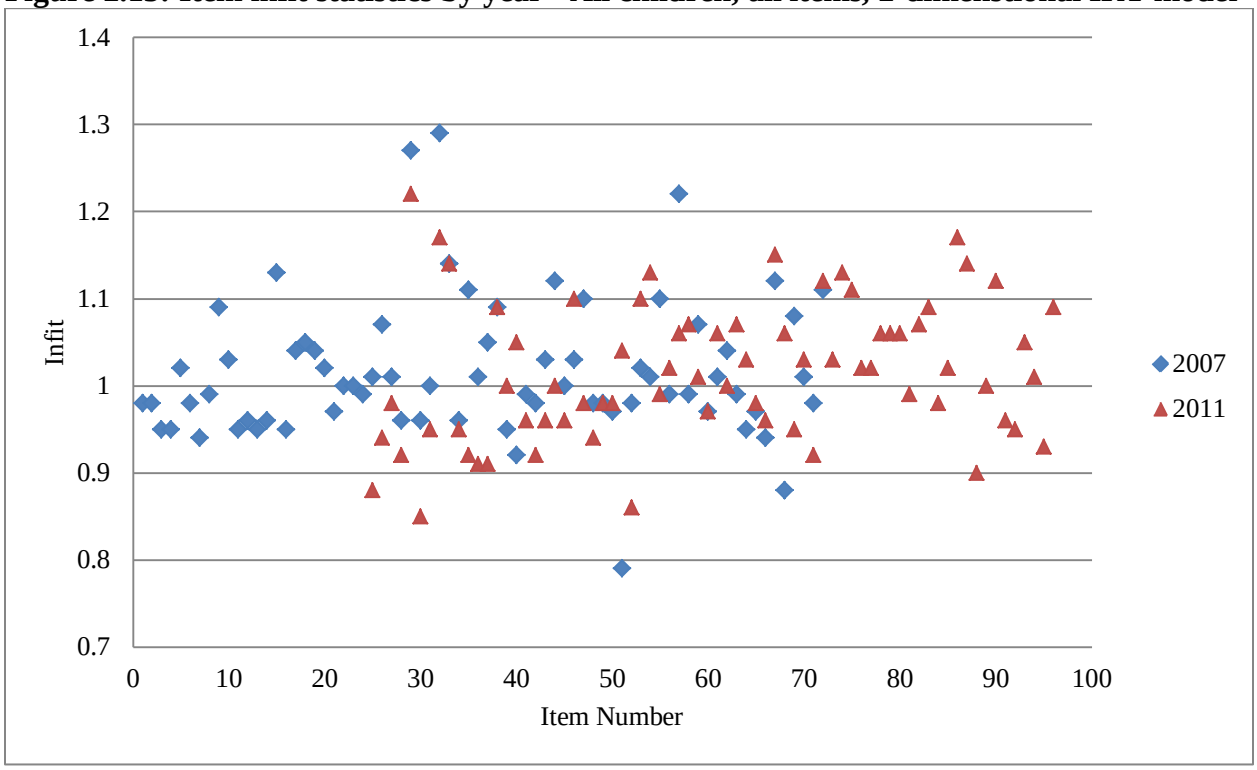


Table 2.4: Item statistics from 2007 for the unidimensional and multidimensional models

	2007: Unidimensional				2007: Multidimensional						
item #	Estimate	SEM	infit	t-stat	Estimate	SEM	infit	t-stat	Anchor?	Gender DIF	Lang DIF
1	-2.765	0.08	0.97	-0.5	-2.954	0.082	0.98	-0.4	no	no	yes
2	-4.14	0.129	0.97	-0.2	-4.326	0.141	0.98	-0.1	no	no	no
3	-0.335	0.06	0.96	-2	-0.532	0.061	0.95	-2.7	no	no	no
4	-2.132	0.067	0.96	-1.1	-2.324	0.069	0.95	-1.5	no	no	no
5	-0.606	0.058	1.02	1.4	-0.802	0.059	1.02	1.4	no	no	no
6	-1.28	0.059	0.97	-1.5	-1.477	0.06	0.98	-1	no	no	no
7	-2.057	0.066	0.94	-1.8	-2.249	0.068	0.94	-1.8	no	no	yes
8	-2.214	0.069	0.97	-0.9	-2.405	0.07	0.99	-0.3	no	no	no
9	-0.056	0.062	1.1	4.1	-0.255	0.063	1.09	3.7	no	no	no
10	-1.924	0.064	1.03	0.9	-2.115	0.066	1.03	0.9	no	no	no
11	-0.397	0.059	0.96	-2.2	-0.595	0.06	0.95	-2.5	no	no	yes
12	0.543	0.07	0.96	-0.9	0.342	0.071	0.96	-1.1	no	no	no
13	0.717	0.075	0.96	-1	0.513	0.077	0.95	-1.1	no	no	no
14	0.095	0.066	0.96	-1.5	-0.108	0.067	0.96	-1.4	no	no	no
15	-0.037	0.064	1.13	5.2	-0.24	0.066	1.13	5.1	no	no	no
16	0.108	0.066	0.96	-1.6	-0.095	0.067	0.95	-1.7	no	no	no
17	-1.136	0.062	1.03	1.8	-1.336	0.063	1.04	2	no	no	no
18	1.104	0.084	1.06	1	0.899	0.087	1.05	0.9	no	no	no
19	-0.101	0.064	1.05	2.1	-0.304	0.065	1.04	1.7	no	no	no
20	0.671	0.074	1.01	0.4	0.466	0.076	1.02	0.5	no	no	no
21	-1.302	0.062	0.97	-1.3	-1.502	0.064	0.97	-1.7	no	no	no
22	-0.295	0.062	1	0.1	-0.497	0.063	1	0.1	no	no	no
23	-0.172	0.063	1.01	0.3	-0.375	0.064	1	0.1	no	no	no
24	0.162	0.066	0.99	-0.2	-0.041	0.068	0.99	-0.4	no	no	no
25	0.651	0.11	0.98	-0.3	0.453	0.117	1.01	0.1	no	no	no
26	-0.165	0.1	0.98	-1	-0.047		1.07	2.1	yes	yes	no
27	-1.617	0.114	1.02	0.3	-1.808	0.122	1.01	0.2	no	no	no
28	0.624	0.109	0.95	-1.1	0.426	0.116	0.96	-0.8	no	no	no
29	1.998	0.155	1.03	0.2	2.033		1.27	1.6	yes	no	no
30	-0.22	0.1	0.92	-3.5	-0.218		0.96	-1.6	yes	no	no
31	0.89	0.115	1.02	0.3	0.643		1	0	yes	no	no
32	0.714	0.111	1.02	0.4	0.904		1.29	3.4	yes	no	yes
33	0.686	0.111	1.02	0.4	0.687		1.14	2.1	yes	no	no
34	-1.284	0.107	0.97	-0.5	-1.475	0.114	0.96	-0.9	no	no	no
35	1.083	0.12	1.01	0.2	1.016		1.11	1.3	yes	no	no
36	-0.546	0.1	1.01	0.3	-0.761		1.01	0.5	yes	no	no
37	-0.788	0.144	1	0.1	-1.178		1.05	0.7	yes	no	yes
38	2.044	0.192	1.08	0.5	1.813	0.24	1.09	0.5	no	no	yes

39	0.578	0.144	1.01	0.2	0.037		0.95	-1.2	yes	no	no
40	0.968	0.152	1.03	0.4	0.314		0.92	-1.5	yes	no	no
41	-0.153	0.138	0.98	-0.7	-0.381	0.153	0.99	-0.4	no	no	no
42	0.252	0.14	0.97	-0.9	-0.402		0.98	-0.7	yes	no	no
43	0.399	0.142	1.01	0.3	0.17	0.158	1.03	0.7	no	no	yes
44	0.325	0.141	1.03	0.7	0.323		1.12	2.1	yes	no	yes
45	0.845	0.15	0.99	-0.1	0.616	0.169	1	0	no	no	no
46	1.12	0.157	1.03	0.3	0.889	0.18	1.03	0.3	no	no	no
47	-0.698	0.143	0.99	-0.2	-1.223		1.1	1.5	yes	no	no
48	0.052	0.139	0.99	-0.4	-0.176	0.154	0.98	-0.7	no	no	no
49	0.882	0.187	1.04	0.5	0.432		0.98	-0.2	yes	no	no
50	0.579	0.181	1.03	0.4	0.165		0.97	-0.6	yes	no	no
51	-1.601	0.213	0.93	-0.3	-1.598		0.79	-1.5	yes	yes	no
52	0.156	0.178	0.99	-0.3	-0.101	0.215	0.98	-0.5	no	no	no
53	1.103	0.192	1.03	0.3	0.846	0.24	1.02	0.2	no	no	no
54	1.478	0.203	1.07	0.5	1.163		1.01	0.1	yes	yes	no
55	-0.309	0.18	0.98	-0.4	-0.909		1.1	1.2	yes	no	yes
56	0.389	0.179	0.97	-0.7	-0.363		0.99	-0.2	yes	no	no
57	1.283	0.197	1.12	0.9	1.211		1.22	1.4	yes	no	no
58	0.436	0.18	1	-0.1	0.179	0.218	0.99	-0.2	no	no	no
59	0.065	0.178	0.97	-0.8	-0.663		1.07	1.2	yes	no	yes
60	-0.261	0.18	0.95	-1	-0.564		0.97	-0.7	yes	yes	no
61	-0.162	0.206	0.98	-0.3	-0.068		1.01	0.3	yes	no	yes
62	0.626	0.207	1.04	0.6	0.373	0.271	1.04	0.6	no	no	no
63	1.885	0.237	1.13	0.6	1.459		0.99	0	yes	no	yes
64	0.93	0.211	1.05	0.5	0.11		0.95	-1.1	yes	yes	yes
65	0.339	0.205	0.97	-0.5	-0.009		0.97	-0.7	yes	no	no
66	0.93	0.211	1.02	0.2	0.453		0.94	-0.8	yes	no	yes
67	1.549	0.226	1.13	0.7	1.293	0.319	1.12	0.7	no	no	no
68	-1.839	0.251	0.93	-0.2	-1.996		0.88	-0.4	yes	no	yes
69	-0.311	0.208	0.95	-0.7	-0.952		1.08	0.7	yes	no	no
70	0.481	0.205	1.02	0.4	0.228	0.268	1.01	0.2	no	no	no
71	-0.387	0.209	0.98	-0.2	-0.638	0.277	0.98	-0.2	no	no	no
72	1.546	1.216	1.12	0.7	1.293	0.319	1.11	0.6	no	no	no

Table 2.5: Item statistics from 2011 for the unidimensional and multidimensional models

	2011: Unidimensional				2011: Multidimensional						
item #	Estimate	SEM	infit	t-stat	Estimate	SEM	infit	t-stat	Anchor?	Gender DIF	Lang DIF
25	-1.005	0.049	0.88	-5.8	-0.749	0.062	0.88	-5.8	no	no	no
26	0.054	0.048	0.94	-2.9	-0.047		0.94	-3.9	yes	no	no
27	-2.687	0.061	0.98	-0.2	-2.433	0.096	0.98	-0.3	no	no	yes
28	-0.744	0.048	0.93	-3.9	-0.488	0.06	0.92	-4.5	no	no	yes
29	1.529	0.058	1.02	0.4	2.033		1.22	3.2	yes	no	no
30	-0.19	0.048	0.84	-10.3	-0.218		0.85	-10	yes	no	yes
31	0.41	0.05	0.95	-1.9	0.643		0.95	-1.9	yes	no	no
32	0.44	0.05	1.08	2.9	0.904		1.17	5.3	yes	no	yes
33	0.354	0.049	1.11	4.5	0.687		1.14	5.2	yes	no	yes
34	-2.178	0.057	0.95	-0.9	-1.924	0.08	0.95	-1.1	no	no	yes
35	0.871	0.052	0.97	-0.9	1.016		0.92	-2.6	yes	no	no
36	-1.088	0.049	0.92	-3.5	-0.761		0.91	-4.5	yes	no	yes
37	-1.579	0.052	0.97	-0.9	-1.178		0.91	-3.3	yes	no	no
38	1.18	0.055	1.08	1.8	1.44	0.075	1.09	2	no	no	yes
39	-0.468	0.048	0.98	-1.2	0.037		1	0.1	yes	no	no
40	-0.236	0.048	0.99	-0.6	0.314		1.05	2.4	yes	no	no
41	-1.092	0.049	0.96	-1.6	-0.836	0.063	0.96	-1.8	no	no	no
42	-0.705	0.048	0.92	-4.6	-0.402		0.92	-4.8	yes	no	no
43	0.526	0.05	0.96	-1.5	0.784	0.065	0.96	-1.4	no	no	no
44	0.018	0.048	0.98	-1.2	0.323		1	-0.2	yes	no	no
45	0.809	0.052	0.96	-1.3	1.068	0.069	0.96	-1.2	no	no	yes
46	0.937	0.053	1.1	2.5	1.196	0.071	1.1	2.6	no	no	no
47	-1.323	0.05	0.91	-3.3	-1.223		0.98	-0.8	yes	no	yes
48	-1.348	0.05	0.94	-2	-1.093	0.065	0.94	-2.2	no	no	no
49	0.131	0.049	0.97	-1.3	0.432		0.98	-0.9	yes	no	yes
50	-0.167	0.048	0.96	-2.2	0.165		0.98	-1.2	yes	no	no
51	-1.612	0.052	0.91	-2.6	-1.598		1.04	1	yes	no	yes
52	-0.9	0.049	0.85	-7.8	-0.645	0.061	0.86	-7.8	no	no	yes
53	1.31	0.056	1.08	1.7	1.569	0.078	1.1	2	no	no	yes
54	0.751	0.052	1.05	1.5	1.163		1.13	3.5	yes	no	no
55	-1.096	0.049	0.97	-1.5	-0.909		0.99	-0.6	yes	no	no
56	-0.493	0.048	1.01	0.8	-0.363		1.02	1.2	yes	no	no
57	0.945	0.053	1.05	1.3	1.211		1.06	1.4	yes	no	yes
58	1.117	0.054	1.06	1.5	1.376	0.074	1.07	1.6	no	no	no
59	-0.669	0.048	0.96	-2.6	-0.663		1.01	0.5	yes	no	no
60	-0.687	0.048	0.95	-2.9	-0.564		0.97	-1.5	yes	no	yes
61	-0.389	0.048	1.06	3.8	-0.068		1.06	3.8	yes	no	yes
62	0.473	0.05	0.99	-0.3	0.731	0.064	1	-0.1	no	no	no

63	0.995	0.053	0.95	-1.2	1.459		1.07	1.6	yes	no	no
64	-0.406	0.048	1	-0.3	0.11		1.03	1.6	yes	no	no
65	-0.318	0.048	0.98	-1.5	-0.009		0.98	-1.3	yes	no	no
66	0.141	0.049	0.94	-2.9	0.453		0.96	-2.1	yes	no	no
67	0.281	0.049	1.14	6	0.538	0.063	1.15	6.3	no	no	yes
68	-2.073	0.056	0.94	-1.3	-1.996		1.06	1.1	yes	no	yes
69	-0.916	0.049	0.86	-7.5	-0.952		0.95	-2.1	yes	no	no
70	0.859	0.052	1.01	0.3	1.118	0.069	1.03	0.8	no	no	no
71	-1.81	0.053	0.93	-1.9	-1.555	0.072	0.92	-1.9	no	no	yes
72	1.478	0.057	1.1	1.8	1.738	0.082	1.12	2.2	no	no	yes
73	0.732	0.051	1.02	0.7	0.991	0.068	1.03	0.9	no	no	no
74	0.326	0.049	1.13	5.2	0.584	0.063	1.13	5.3	no	no	no
75	0.927	0.053	1.1	2.6	1.186	0.07	1.11	2.8	no	no	yes
76	1.229	0.055	1.01	0.2	1.489	0.076	1.02	0.5	no	no	no
77	0.59	0.051	1.02	0.8	0.848	0.066	1.02	0.8	no	no	no
78	-0.142	0.048	1.05	3	0.115	0.06	1.06	3.3	no	no	no
79	0.302	0.049	1.05	2.3	0.56	0.063	1.06	2.6	no	no	no
80	1.682	0.059	1.05	0.8	1.943	0.088	1.06	1	no	no	no
81	-0.327	0.048	0.98	-1.1	-0.071	0.06	0.99	-0.8	no	no	no
82	1.37	0.056	1.07	1.3	1.63	0.079	1.07	1.4	no	no	no
83	2.079	0.063	1.08	1	2.341	0.101	1.09	1.1	no	no	yes
84	-0.986	0.049	0.98	-0.8	-0.731	0.062	0.98	-0.9	no	no	no
85	0.22	0.049	1.01	0.6	0.477	0.062	1.02	0.7	no	no	no
86	-0.006	0.048	1.16	8.2	0.252	0.061	1.17	8.3	no	no	yes
87	0.233	0.049	1.13	5.8	0.49	0.062	1.14	6	no	no	yes
88	-0.832	0.048	0.9	-5.7	-0.576	0.061	0.9	-5.6	no	no	no
89	0.587	0.051	0.99	-0.3	0.845	0.066	1	-0.1	no	no	no
90	1.012	0.053	1.11	2.6	1.271	0.072	1.12	3	no	no	yes
91	-0.377	0.048	0.95	-2.8	-0.12	0.06	0.96	-2.2	no	no	no
92	-0.053	0.048	0.94	-3.2	0.204	0.061	0.95	-2.9	no	no	no
93	-0.549	0.048	1.05	2.8	-0.293	0.06	1.05	3	no	no	yes
94	1.424	0.057	1.01	0.1	1.684	0.081	1.01	0.3	no	no	no
95	-0.259	0.048	0.93	-4.3	-0.002	0.06	0.93	-4.1	no	no	no
96	1.386	0.43	1.07	1.4	1.645	0.08	1.09	1.7	no	no	no

A2: R Code for Simulations

R Code for simulations 1-6:

```
#------------------------
# In all, I assume W1 is not affected by A, and exclude observed exogenous variables, V
#------------------------
set.seed(100)
n <- 100000
C<-rnorm(n,0,4)
W0<-rnorm(n,0,4)
#------------------------
# Run 1: Example for figure 3.1: estimand I, controlling for Y0
# No unmeasured confounding C
Y0<-rnorm(n,0.5*W0,4)
A<-rbinom(n,1,1/(1+exp(-0.5*W0-0.5*Y0)))
W1<-rnorm(n,W0+Y0,4)
Y1<-rnorm(n,W0+2*Y0+A+W1,4)
est1 <- glm(Y1~A+W0+W1+Y0)
#-----------------------------------------
# Run 2: Example for figure 3.3: estimand I
# Introduce unmeasured confounder C that affects Y(0), Y(1) and A
Y0<-rnorm(n,0.5*W0+C,4)
A<-rbinom(n,1,1/(1+exp(-0.5*W0-0.5*Y0-0.5*C)))
W1<-rnorm(n,W0+Y0,4)
Y1<-rnorm(n,W0+Y0+A+W1+C,4)
est2 <- glm(Y1~A+W0+W1+Y0)
#-----------------------------------------
# Run 3: Example for figure 3.5: estimand II, not controlling for Y(0)
# Unmeasured confounder C
Y0<-rnorm(n,0.5*W0+C,4)
A<-rbinom(n,1,1/(1+exp(-0.5*W0-0.5*Y0-0.5*C)))
W1<-rnorm(n,W0+Y0,4)
Y1<-rnorm(n,W0+Y0+A+W1+C,4)
Yd<-Y1-Y0
est3 <- glm(Yd~A+W0+W1)
#-----------------------------------------
# Run 4: Example for figure 3.6: estimand II, not controlling for Y(0)
# Confounder C, assume Y(0) does not affect A, W(1), or Y(1); i.e., no confounding by Y(0)
Y0<-rnorm(n,0.5*W0+C,4)
A<-rbinom(n,1,1/(1+exp(-0.5*W0-0.5*C)))
W1<-rnorm(n,W0,4)
Y1<-rnorm(n,W0+A+W1+C,4)
Yd<-Y1-Y0
est4 <- glm(Yd~A+W0+W1)
#-----------------------------------------
# Run 5: Example for figure 3.7: estimand III
```

```
# Confounder C, assume Y(0) does not affect A, W(1), or Y(1); i.e., no confounding by Y(0)
# Assumption (11) but W(0) affects A and Y(1)
Y0<-rnorm(n,0.5*W0+C,4)
A<-rbinom(n,1,1/(1+exp(-0.5*W0-0.5*C)))
W1<-rnorm(n,W0,4)
Y1<-rnorm(n,W0+A+W1+C,4)
# Reshape wide to long
id <- paste("id", 1:n, sep="")
data_wide <- data.frame(id,C,A,W0,Y0,W1,Y1)
data_long <- reshape(data_wide,
        varying = 4:7,
        idvar = "id",
        direction = "long",
        timevar = "T",
        new.row.names = NULL,
        sep = "")
est5 <- glm(Y~A+W+T+A*T,data=data_long)
#----------------------------------------
# Run 6: Example for figure 3.8: estimand III
# Confounder C, assume Y(0) does not affect A, W(1), or Y(1); i.e., no confounding by Y(0)
# Assumption (11) and W(0) does not affect A or Y(1)
Y0<-rnorm(n,0.5*W0+C,4)
A<-rbinom(n,1,1/(1+exp(-0.5*C)))
W1<-rnorm(n,W0,4)
Y1<-rnorm(n,A+W1+C,4)
# Reshape wide to long
id <- paste("id", 1:n, sep="")
data_wide <- data.frame(id,C,A,W0,Y0,W1,Y1)
data_long <- reshape(data_wide,
        varying = 4:7,
        idvar = "id",
        direction = "long",
        timevar = "T",
        new.row.names = NULL,
        sep = "")
est6 <- glm(Y~A+W+T+A*T,data=data_long)
#----------------------------------------
# Run 7: Example adding Y(0) affects A into run 4
Y0<-rnorm(n,0.5*W0+C,4)
A<-rbinom(n,1,1/(1+exp(-0.5*W0-0.5*Y0-0.5*C)))
W1<-rnorm(n,W0,4)
Y1<-rnorm(n,W0+A+W1+C,4)
Yd<-Y1-Y0
est7 <- glm(Yd~A+W0+W1)
#----------------------------------------
# Run 8: Example adding Y(0) affects A into run 6
```

```
Y0<-rnorm(n,0.5*W0+C,4)
A<-rbinom(n,1,1/(1+exp(-0.5*Y0-0.5*C)))
W1<-rnorm(n,W0,4)
Y1<-rnorm(n,A+W1+C,4)
# Reshape wide to long
id <- paste("id", 1:n, sep="")
data_wide <- data.frame(id,C,A,W0,Y0,W1,Y1)
data_long <- reshape(data_wide,
        varying = 4:7,
        idvar = "id",
        direction = "long",
        timevar = "T",
        new.row.names = NULL,
        sep = "")
est8 <- glm(Y~A+W+T+A*T,data=data_long)
#------
est_all <-rbind(est1$coeff["A"],est2$coeff["A"],est3$coeff["A"],est4$coeff["A"],
est5$coeff["A:T"],est6$coeff["A:T"],est7$coeff["A"],est8$coeff["A:T"])
est_all
```

A3: Supplementary Information for Estimation

SuperLearner

SuperLearner (SL)[102] is a non-parametric, machine-learning tool that "learns" from the observed data by using a candidate set of algorithms (or estimators) and a pre-specified loss function that assigns a measure of performance to each of the algorithms. Briefly, there are three key components to SL:

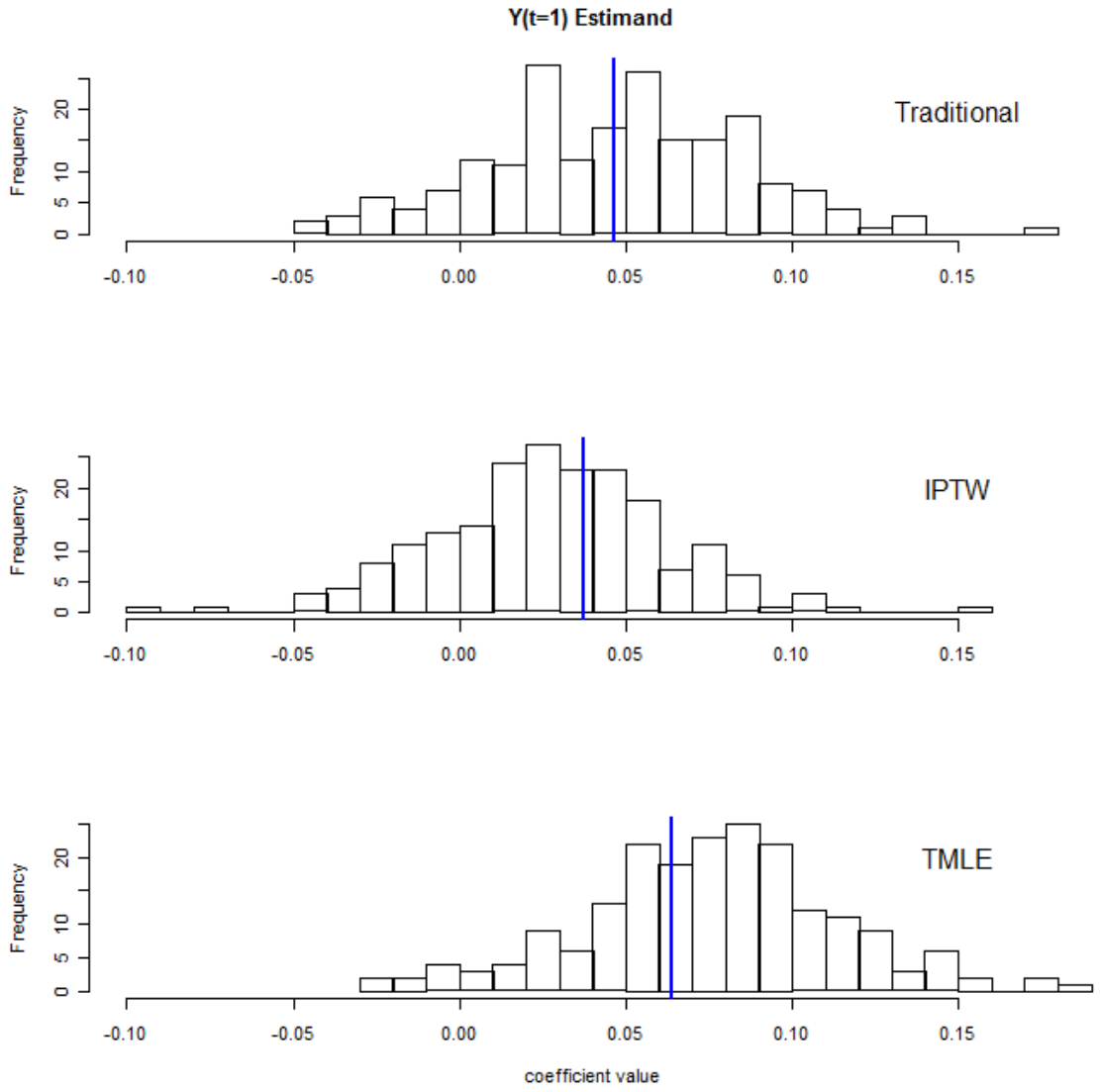
1) SL uses a library of algorithms for prediction. The algorithms can be diverse, simple (i.e., logistic regression), complex (i.e., neural nets), numerous, and can include user defined algorithms.

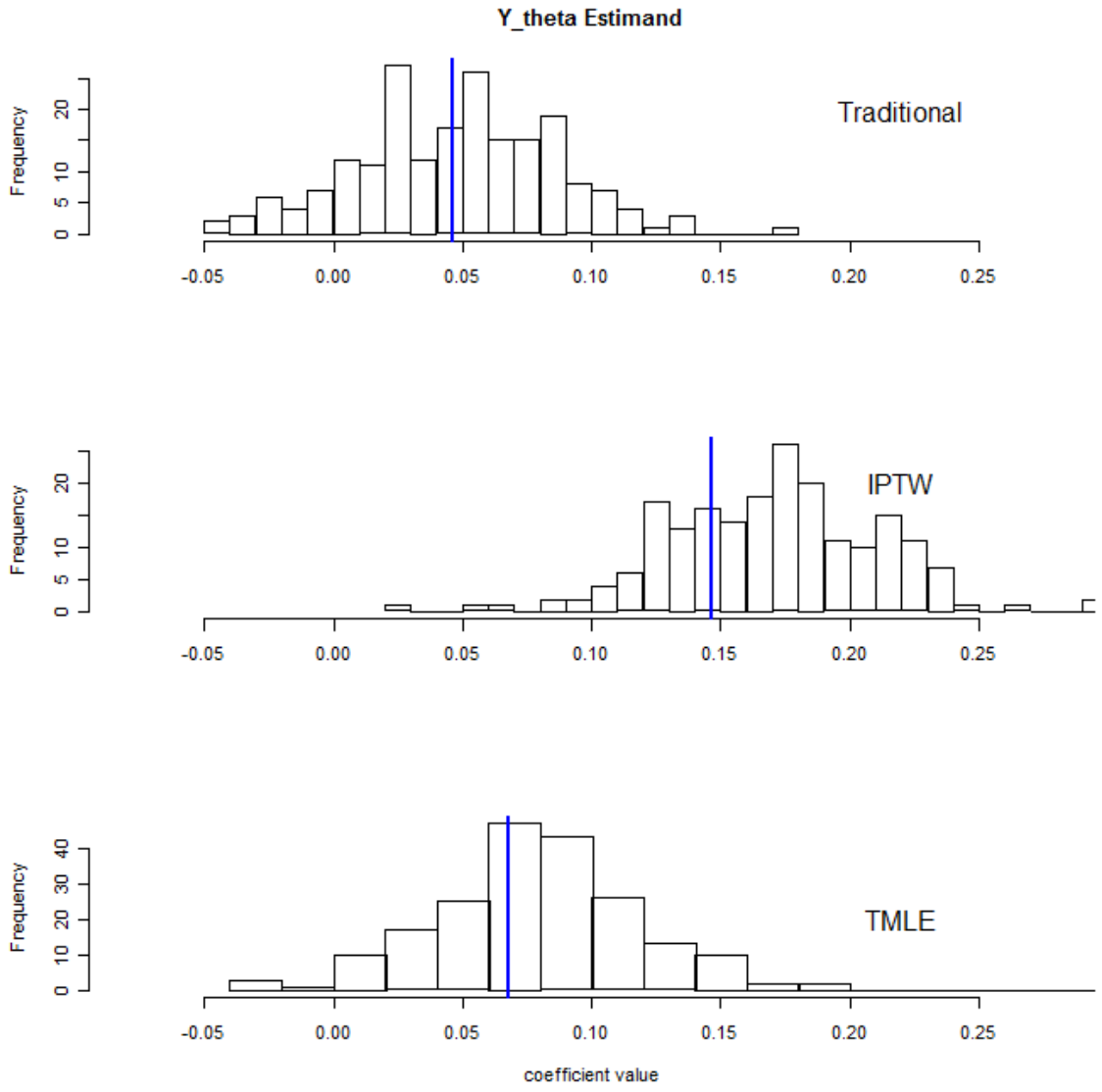
2) The predictive performance of each algorithm is assessed using V-fold cross-validation. Cross-validation involves partitioning the sample into a user-specified number of training and validation sets. A training set is used to construct the candidate estimators (i.e., fit the regression) and the corresponding validation set is then used to assess the performance (i.e., estimate the risk) of the candidate algorithms. The validation set rotates by the number of partitions such that each set is used as the validation set once. Risk is defined using a loss function, for example, if we use the squared error loss function then our estimate of the risk corresponds to the estimated mean squared error loss on the validation sets. The "best" algorithms typically have the smallest empirical risk averaged over all the validation sets.

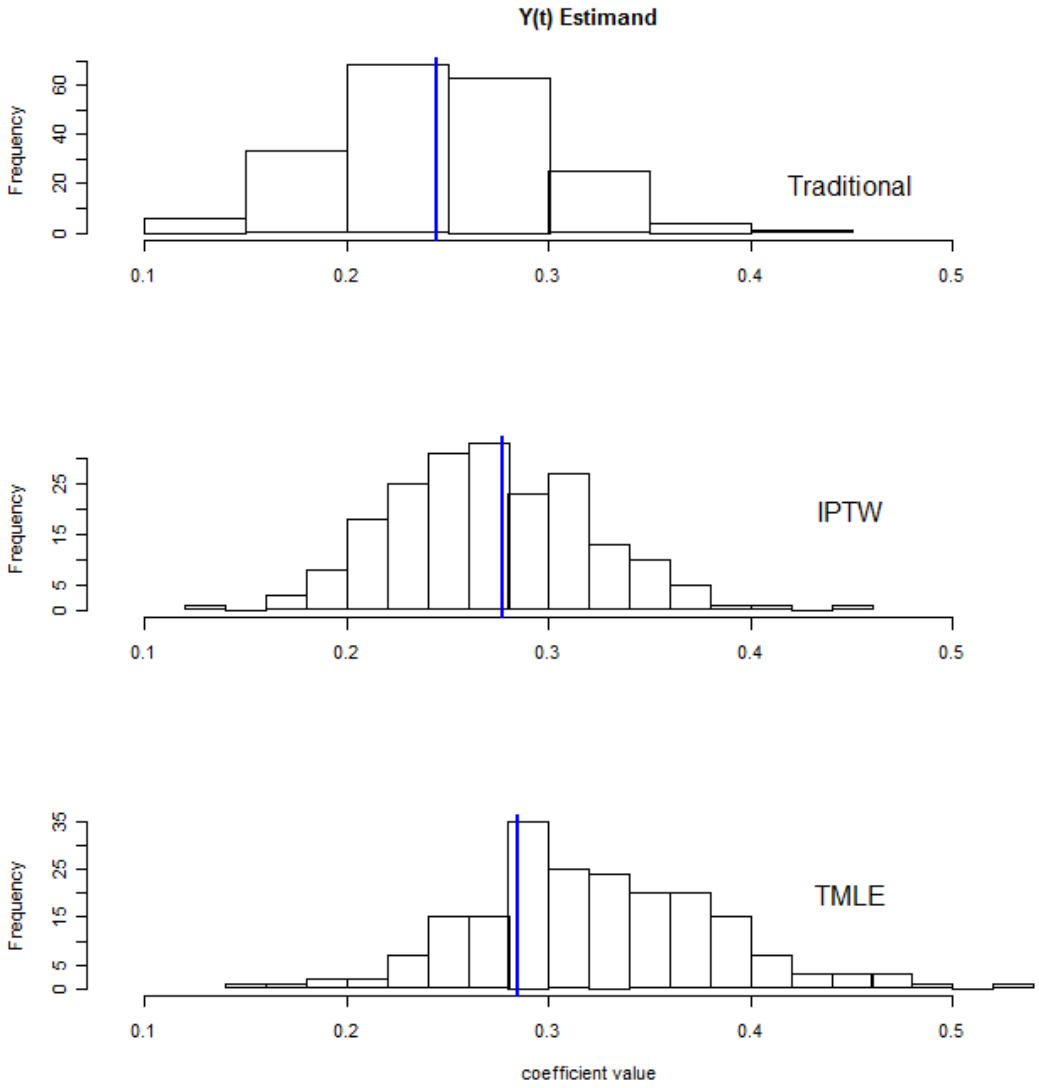
3) The library of algorithms is augmented with new algorithms, which are weighted averages of the algorithms from the previous step. The weighted algorithm with the smallest cross-validated risk is the "super learner" estimator and is expected to outperform any single algorithm. If the true model is included in the library of algorithms, then SL will do as well as the true model. (Note that we can include a parametric model in the SuperLearner library.)

Bootstrap Estimates

Figures 4.4a-c: Distributions of coefficients for Estimands I, II, & III on 200 bootstrap samples




Y_theta Estimand
Traditional
IPTW
TMLE
Frequency
0
5
10
20
30
40
-0.05
0.00
0.05
0.10
0.15
0.20
0.25
coefficient value


Y(t) Estimand
Frequency
0
20
40
60
0.1
0.2
0.3
0.4
0.5
Traditional
Frequency
0
5
15
25
IPTW
Frequency
0
5
15
25
35
TMLE
coefficient value