Strengthening Causal Inferences: Examining Instrument-Free Approaches to Addressing Endogeneity Bias in the Evaluation of an Integrated Student Support Program

Author: Jordan L. Lawson

Boston College Electronic Thesis or Dissertation, 2019

Boston College
Lynch School of Education

Department of
Measurement, Evaluation, Statistics, and Assessment

STRENGTHENING CAUSAL INFERENCES: EXAMINING INSTRUMENT-FREE APPROACHES TO ADDRESSING ENDOGENEITY BIAS IN THE EVALUATION OF AN INTEGRATED STUDENT SUPPORT PROGRAM

Dissertation
by

JORDAN L. LAWSON

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

AUGUST 2019

STRENGTHENING CAUSAL INFERENCES: EXAMINING INSTRUMENT-FREE APPROACHES TO ADDRESSING ENDOGENEITY BIAS IN THE EVALUATION OF AN INTEGRATED STUDENT SUPPORT PROGRAM

Jordan L. Lawson, Author

Laura M. O'Dwyer, Chair

Education researchers are frequently interested in examining the causal impact of academic services and interventions; however, it is often not feasible to randomly assign study elements to treatment conditions in the field of education (Adelson, 2013). When assignment to treatment conditions is non-random, the omission of any variables relevant to treatment selection creates a correlation between the treatment variable and the error in regression models. This is termed endogeneity (Ebbes, 2004). In the presence of endogeneity, treatment effect estimates from traditionally used regression approaches may be biased.

The purpose of this study was to investigate the causal impact of an integrated student support model, namely City Connects, on student academic achievement. Given that students are not randomly assigned to the City Connects intervention, endogeneity bias may be present. To address this issue, two novel and underused statistical approaches were used with school admissions lottery data, namely Gaussian copula regression developed by Park and Gupta (2012), and Latent Instrumental Variable (LIV) regression developed by Peter Ebbes (2004). The use of real-world school admissions

lottery data allowed the first-ever comparison of the two proposed methods with Instrumental Variable (IV) regression under a large-scale randomized control (RCT) trial. Additionally, the researcher used simulation data to investigate both the performance and boundaries of the two proposed methods compared with that of OLS and IV regression.

Simulation study findings suggest that both Gaussian copula and LIV regression are useful approaches for addressing endogeneity bias across a range of research conditions. Furthermore, simulation findings suggest that the two proposed methods have important differences in their set of identifying assumptions, and that some assumptions are more crucial than others.

Results from the application of the Gaussian copula and LIV regression in the City Connects school lottery admissions study demonstrated that receiving the City Connects model of integrated student support during elementary school has a positive impact on mathematics achievement. Such findings underscore the importance of addressing out-of-school barriers to learning.

DEDICATION

To my son, Isaiah Alejandro-Levonne Lawson. You taught me strength and true love. You'll always be my real life hero and this one is for you.

To my cherished friend, Eric S. Strong. I am forever appreciative of your kind heart, character, and unconditional support throughout the years. I'll miss you and this is also for you. Fly high.

ACKNOWLEDGMENTS

First and foremost, I want to sincerely thank my dissertation chair, Dr. Laura O'Dwyer, for her support, guidance, and kindness along the way. Without her mentorship, this work would not have been completed and, as such, I am forever grateful that I was lucky enough to have been her advisee. Thank you for not only being an intelligent, thoughtful academic but also a great person.

Additionally, I want to thank my dissertation committee members. Dr. Eric Dearing, thank you for always believing in me and engaging me in such interesting and thoughtful discussions about statistics. Talking and brainstorming with you has helped me immensely. Dr. Walsh, thank you for the opportunity to pursue this work and for helping instill in me the importance of using statistics for social good. My time at City Connects has been an invaluable learning experience. Dr. Michael Russell, thank you for your encouraging words and critical eye when reviewing my work. Your support along the way meant a lot.

To the Lawson family, thank you for your love and support along the way. You taught me values, hard work, faith, tough love, and provided me with the motivation I needed to finish. You all had a piece in helping me get to this point, and I love you.

Lastly, thank you to the Berry's for reminding me that great people still exist.

Table of Contents

CHAPTER 1: INTRODUCTION

Overview of the Problem

In many fields, experimental research designs remain the gold standard for isolating the effects of a treatment or intervention (Shadish, Cook, & Campbell, 2001; Keppel & Wickens, 2004). When conducting an experimental study, researchers assign study elements (i.e., individuals or clusters of individuals) to treatment and control groups through a random mechanism, which aims to ensure group comparability on all observed and unobserved characteristics (Alemayehu, Alvir, Jones, & Willke, 2011). When exposure to the treatment is allocated randomly, unrecognized confounding effects become statistically unlikely, provided that the sample size is sufficient (Alemayehu et al., 2011). Furthermore, because randomization ensures that the groups are comparable on unobserved and observed confounders, the notion of exchangeability – i.e., switching which group receives treatment does not change the average outcome – can be invoked and thus causation can be inferred (Hernan & Robins, 2018). As a result, experimental designs allow one to reasonably argue that the observed difference between the treatment and control groups on the outcome is attributable to the intervention.

However, it is often the case in educational research that subjects cannot be assigned randomly to conditions; this may be due to lack of feasibility or, even more plausibly, ethical concerns (Adelson, 2013). Adelson among others, points out that factors such as school structures, student needs, and economic constraints limit the possibility of randomly assigning students to a particular condition (2013). Thus, educational researchers who aim to make causal inferences about the effect of some treatment or program must often contend with data that are collected as part of a study in which units are assigned to the treatment through some form of non-random assignment process. Non-random assignment, however, raises the possibility of threats to the internal validity of a study, of which selection bias threats may be the most challenging for making causal inferences (Kaplan, 2009; Keppel & Wickens, 2004). Even with the challenges that non-random assignment procedures often create, researchers remain interested in examining if, and to what degree, an intervention impacts a student outcome.

Despite questions in educational research often being causal in nature, the non-random assignment of elements to treatment conditions often preclude credible causal inferences (Hernan & Robins, 2013). Technically, the central issue is that causal questions require that the variation in the treatment be exogenous (Gerring, 2011; Pearl, 2009). In this particular context, exogenous means that the assignment of study elements to the treatment and control conditions does not depend on the outcome variable being studied, or on any variables related to the outcome variable (Stock & Watson, 2014). In the case of research scenarios with non-random assignment, however, treatment assignment may no longer be independent of the outcome variable or of its correlates. Moreover, when confounding variables are related to both treatment assignment and the outcome variable, the types of statistical analyses typically conducted to estimate a treatment effect (e.g., analysis of covariance, ordinary least squares regression) may result in misleading findings. Specifically, when treatment assignment is not independent of the outcome variable or of its correlates, some of the assumptions of the statistical models may be violated, possibly leading to incorrect inferences (Keppel & Wickens, 2004).

As will be elaborated on in the remainder of this chapter and throughout this dissertation, making causal inferences in the absence of random assignment remains an intractable problem and researchers continue to explore statistical adjustments to addressing the lack of exogeneity. The instrumental variable (IV) regression method is one such approach that shows promise for addressing this issue (Ebbes, 2004; Hueter, 2016). Under the traditional IV approach, an ordinary least squares regression model is augmented by the introduction of an instrumental variable, which is included to partition out the problematic correlation between the assignment variable and the structural error term (Ebbes, 2004). However, the instrumental variable approach has some limitations when appropriate observed instrumental variables cannot be identified (Ebbes, 2004; Hueter, 2016).

In an effort to address this challenge, Ebbes (2004) and Park and Gupta (2012) have proposed a set of procedures that do not require researchers to identify observed instruments. Referred to as instrument-free methods, the Latent Instrumental Variable (LIV) approach developed by Peter Ebbes (2004) and the Gaussian copula approach developed by Park and Gupta (2012), show immense promise for addressing the challenge of identifying an appropriate instrumental variable that supports causal inferences. However, a review of the extant literature in this area reveals that their empirical application has been limited, and neither approach has been adopted for use in educational research. Moreover, comparisons among the Gaussian copula approach, the LIV approach, and the traditional IV approach are scarce, and a comparison using real-world school lottery data to evaluate the impacts of an integrated student support program has never been conducted.

In response to the dearth of research in this area, the goal of this dissertation research is to evaluate the efficacy of an integrated student support, namely City

Connects, for improving student academic achievement. Given that City Connects does not randomly assign students to receive the intervention, any evaluation of the City Connects intervention must contend with the possibility of endogeneity selection bias. To address this issue, the researcher investigated the utility of two novel and underused statistical approaches for dealing with endogeneity bias, namely Gaussian copula and Latent Instrumental Variable (LIV) regression. The methods were investigated in a two-step manner. First, using simulation data, the researcher investigated both the performance and boundaries of the two methods compared with that of OLS and Instrumental Variable (IV) regression. Subsequently, the researcher applied the two methods to real-world data compiled from a large-scale lottery study that is being conducted to evaluate the effects of the City Connects intervention. To situate the relevance of this dissertation research, the sections that follow provide a description of the City Connects intervention followed by a discussion of the broader context for making causal inferences, outlining the inferential framework upon which this dissertation work rests and the logic invoked. Subsequently, the traditional IV approach is presented along with the methodological extensions that are at the heart of this dissertation research. This chapter ends with a discussion of the significance of this research. Note that this dissertation research frames non-randomization as an obstacle to sound educational evaluation and is built on the premise that causal inference research is one of several worthy forms of educational research for examining the effects of policy, programs, and interventions.

City Connects Intervention

Prior research has shown that out-of-school factors can have significant impacts on students' readiness to learn and thrive in school, accounting for up to two-thirds of the variance in academic achievement (Rothstein, 2010). The academic impact of out-of-school factors is especially notable for students growing up in poverty. In fact, when compared to students from middle-class and affluent families during early school years, children in poverty score, on average, one-half to one full standard deviation lower on tests of achievement (Kaushal, Magnuson, & Waldfogel, 2012). This achievement gap has direct consequences for students' educational attainment and life outcomes; for example, it has been estimated that children growing up in extreme poverty are up to 12 times less likely to graduate from high school than youth from middle-class families (Duncan, Brooks-Gunn, Yeung & Smith, 1998). In response to such a concern, researchers at Boston College developed the City Connects intervention (Walsh et al., 2014).

City Connects is an integrated student support model offering student support in high-poverty, urban schools. At the core of the intervention is a full-time Coordinator, trained as a Masters'-level licensed school counselor or social worker. Every fall, the Coordinator meets with every classroom teacher to identify each student's strengths and needs. Specifically, Coordinators engage in a conversation with classroom teachers using a series of guiding questions aimed at eliciting teacher insights on student strengths and needs across four developmental domains (academic, social/emotional/behavioral, health, and family). During and following this conversation, City Connects Coordinators develop tailored plans for each student, identifying particular enrichment and service programs that best suit the strengths and needs of each individual student (Progress Report, 2018).

To identify these resources, and to allow for the tracking and follow-up of service delivery, Coordinators use a proprietary web-based database designed for the intervention to find specific service providers based on factors such as service type(s), geographical location, schedule, transportation requirements, and family capacity to support participation (e.g., access to insurance). Coordinators then connect students and their families with service providers, coordinate the provision of services, monitor service quality and appropriateness, and maintain partnerships with community providers (Progress Report, 2018). Throughout the entire process, the City Connects Coordinator works closely with students and their families to ensure service delivery.

To date, multiple studies have demonstrated the efficacy of the City Connects intervention. Dearing et al. (2016) investigated the impact of City Connects on the mathematics and reading achievement of first-generation immigrant children living in high poverty, urban contexts. The study revealed significant and practically important positive effects in both mathematics and reading performance during elementary school years. Furthermore, Walsh et al. (2014) reported both higher report card scores and higher performance on middle school English language arts and mathematics tests for students participating in the City Connects intervention. Lee-St. John et al. (2018) investigated the association between participation in City Connects and high school dropout, finding that City Connects students had approximately half the odds of dropout. Although such studies provide compelling evidence for the effectiveness of City Connects for addressing non-academic barriers to learning, all research of the intervention to-date has relied on quasi-experimental design. This dissertation provided the first-ever evidence from a randomized control trial design demonstrating the efficacy of the City Connects intervention.

Role of Causal Inference and Experimentation in Educational Research

Objects and facts in isolation are hardly of interest to scientists; instead, science is an enterprise concerning itself with relationships (Ozen, 2011). The educational sciences are thus no different, with educational researchers seeking to understand the relationship between the implementation of educational programs or policies and student outcomes. Given that education is widely viewed as a mechanism for improving the human condition, it comes as no surprise that federal agencies, policy-makers, researchers, and educators alike are deeply invested in improving educational outcomes. Despite the social consensus around the importance of education, public resources remain limited and education must often contend with the demands and needs of the public for other services, such as health care and public safety (Willett & Murnane, 2010). As a result, decisions regarding the allocation of resources to educational activities must be justified, with arguably this justification coming from empirical evidence regarding the programs and conditions that lead to improved student outcomes. Consequently, researchers have become increasingly interested in answering questions regarding the determinants of student achievement and the effectiveness of programs and policies devoted to improving these outcomes.

There has long been a call for basing educational decisions in strong empirical evidence. In 1913, Paul Hanus, a Harvard Professor and the first dean of the Harvard Graduate School of Education, delivered a speech to the National Education Association (NEA) in which he argued that systematic research must be conducted within the field of education and that the findings from such research ought to guide decision-making (Willett & Murnane, 2010). Specifically, he stated that, "We are no longer disputing whether education has a scientific basis; we are trying to find that basis" (Willett & Murnane, 2010). More than a century later, we note that the U.S. Department of Education's Institute for Educational Sciences (IES) and the organization it sponsors, "The What Works Clearinghouse," has set evidence standards that give focus to scientifically-based educational research, with particular preference given to experimental research that involves random-assignment. Thus, we see that "the basis" of which Hanus spoke in his address to the NEA in 1913 has been identified as experimentation.

The focus of this dissertation research on comparing statistical approaches that support causal inference aligns itself with the particular conception of educational research and evaluation put forth by IES, thereby adopting a successionist framework for thinking about causality and relying on regularity and counterfactual logic as an argumentative basis (Gates & Dyson, 2016). The purpose of this work, however, is not to evaluate the epistemological merits of a particular causal framework; instead, this discussion serves to position the work within the broader context of educational research, providing the necessary context and line of reasoning for the methodological developments that will soon follow.

Causal Inference

Causality as a concept comes naturally to human beings, as we tacitly presuppose causal connections between phenomena on a daily basis in order to successfully navigate the world around us (Faye, 2014; Peterson, 1898). Thus, when faced with causal statements such as, “pressing the power button caused the computer to start,” one is able to easily make intuitive sense of what this means; furthermore, direct experiences of this act happening warrants such a claim in our daily lives (Faye, 2014; Lee-St. John, 2012). Causal claims in science are akin to the causal claims we make in our daily lives, differing only in how they are warranted, as the former demands much stronger, detailed theoretical argument and sophisticated methods of observation (Faye, 2014). The basis for warranting causal claims in science can be found in Hume’s (1748) seminal work, *An Enquiry Concerning Human Understanding,* in which he defined causation as follows:

> We may define a cause to be *an object, followed by another, and where all the objects, similar to the first, are followed by objects similar to the second* [emphasis in original]. Or, in other words, *where, if the first object had not been, the second never had existed.* [emphasis in original] (Peterson, 1898, p. 44.)

The above definition provided by Hume is a noteworthy starting point because it was the first documented definition of causation that invoked *counterfactual* logic. Thus, by relying on such logic, we can see that the previously given example of a causal statement, “pressing the power button caused the computer to start” can be formally explained by the statement, “if the power button had not been pressed, the computer would never had started.” More generally, we can formally explain any causal statement “X causes Y” by its counterfactual statement – “if X had not occurred, then Y would not have occurred” (Lee-St. John, 2012; Hume, 1748; Kaplan, 2009). This counterfactual conditional proposition has remained integral to the theory of causation (Kaplan, 2009).

While Hume’s counterfactual definition of causation has been challenged by many, it was most notably John Stuart Mill who first debated and refined Hume’s analysis of causation (Kaplan, 2009; Lee-St. John, 2012). Mill's main contribution to the theory of causation comes from qualifying Hume's view of invariable succession, the identifying feature of causation made known to us only through experience. Specifically, Mill asserted that causation is a sequence of events (i.e., a succession) that is not only invariable but also *unconditional* (Peterson, 1898). In his *A System of Logic*, Mill (1843) writes:

> If there be any meaning which confessedly belongs to the term necessity, it is *unconditionalness* [emphasis in original]. That which is necessary, that which *must* [emphasis in original] be, means that which will be, whatever supposition we may make in regard to all other things. The succession of day and night evidently is not necessary in this sense…..We may define, therefore, the cause of a phenomenon to be the antecedents, or the concurrence of antecedents, on which it is invariably and *unconditionally* [emphasis in original] consequent. (Peterson, 1898. p. 46)

In plain language, Mill points out that relying on the single criterion of invariable succession for defining a cause inevitably leads to non-causal regularities warranting causal claims, such as reasoning that day causes night and vice versa. This will happen because the two events, given human experience, always accompany one another, even though day is not the cause of night and night does not cause day (Peterson, 1898). As a result, Mill (1748) modifies Hume's definition of causation, adding the qualification that phenomena be unconditionally conjoined (Kaplan, 2009; Peterson, 1898). This is important because hitherto in science and philosophy, we were allowed to arbitrarily pick single objects and label them a cause, essentially ignoring all other conditions equally necessary for an effect to occur (Kaplan, 2009; Hulswit, 2002). Thus, in moving from the work of Hume to Mill, the causal statement changes from, "X causes Y" to, "X causes Y *if and only if* Z is given," where Z is an auxiliary set of true statements consistent with the antecedent, X (Hulswit, 2002; Lee-St. John, 2012). We can view Z as *assumptions* or *premises* (e.g., laws of nature) that, when conjoined with the antecedent, form the necessary set of conditions for an effect to invariably and unconditionally occur (Lee-St. John, 2012). This idea had profound implications, setting the foundation for the types of experimental designs that are considered the gold-standard in research and to which the field of education aspires (Kaplan, 2009).

Drawing from his idea of unconditionalness, Mill argued that causal claims were only warranted when regularities governed by a constant law were isolated from some greater field of circumstances preceding or following that same phenomenon by chance (Huslwitz, 2002: Kaplan, 2009; Lee-St. John, 2012). Namely, Mill's approach was to test for causality by observing the presumed causal connection between phenomena under varying conditions or situations (Hulswitz, 2002: Kaplan, 2009). Additionally, Mill posited three conditions needed for causal inference: 1) the cause must precede the effect (*temporal precedence*); 2) the cause and effect must be related (*covariation*); and 3) alternative possible explanations for the effect must be ruled out (Kaplan, 2009). The third condition was of special interest to Mill, and, as a result, he proposed a set of methods for dealing with this condition. The first method is known as *The Method of Agreement*, which states that the effect will be present when the cause is present. Secondly is *The Method of Differences*, stating that the effect will be absent when the cause is absent; and the third method is known as *The Method of Concomitant Variation*, which states that when both of the first two conditions are observed, causal arguments are strengthened because alternative explanations for the covariation between the cause and effect have been ruled out (Kaplan, 2009). What is important to note is that this idea of varying the conditions in order to rule out alternative explanations gave rise to the notion of experimental manipulation; furthermore, we can see that the idea of a *control group* is implicit within Mill's three methods detailed above, thus setting the foundation for experimental designs (Hulswitz, 2002; Kaplan, 2009).

Experimental Design

Following Mill's treatise on experimental logic came the work of Campbell, which has been highly influential in shaping social scientists' understanding of causality, especially in fields such as education and psychology (Lee St. John, 2012; Kaplan, 2009; Shadish, 2010; West & Thoemmes, 2010). Drawing heavily from Mill, Campbell and Stanley (1963) laid out the logic of experimental and quasi-experimental designs in their seminal work, *Experimental and Quasi-Experimental Designs for Research* (Kaplan, 2009; West & Thoemmes, 2010). In this monograph, they detail the major sources of confounding in research designs and expand upon the notion of *internal validity*, a concept previously created and introduced by Campbell in 1957 (Kaplan, 2009). Specifically, Campbell and Stanley used the following question to characterize internal validity: *Did in fact the experimental treatments make a difference in this specific experimental instance?* Campbell and Stanley (1963) shortly thereafter, stated that internal validity is "the basic minimum without which any experiment is uninterpretable" (pg. 5). Thus, they reintroduce the concept of internal validity as the sine qua non of experimental success (Campbell & Stanley, 1963; Kaplan, 2009). Campbell and Stanley also introduce the idea of *external validity*, which related to the question of generalizability and subsequently state that the ideal experimental design is a design that is strong in both forms of validity (1963). Noteworthy, however, and of direct relevance for this dissertation research, is that the work of Campbell (1957) and Campbell and Stanley (1963) conceptualize internal validity as the degree to which causal claims are warranted. As a result, Campbell (1957), Campbell and Stanley (1963), and Campbell, Cook, and Shadish (2002), along with many other later works, focus on identifying factors that can serve as threats to internal validity and the causal claims made from experimentation (Kaplan, 2009). This focus on identifying extraneous variables that should be controlled is a direct result of Mill's earlier works, viz. the condition of no plausible alternative explanations.

Having established the importance of internal validity for warranting causal claims, Campbell and Stanley (1963) identified eight classes of extraneous variables that may affect the internal validity of a study. Campbell and Stanley's (1963) original threats, presented in the order in which they described them, include *history*, *maturation*, *testing*, *instrumentation*, *statistical regression*, *selection bias*, *experimental mortality*, and *selection maturation interaction* as potential confounds. In light of these threats, building a causal-argument then becomes a two-step process, whereby one first examines the degree to which a given research design and implementation is vulnerable to the aforementioned threats and secondly devises a strategy for systematically addressing the identified threats (Lee-St. John, 2012). Campbell and Stanley focus on strategies for addressing threats to internal validity that are primarily *design-based* (Lee St. John, 2012). In particular, Campbell and Stanley (1963) recommended adding a group that does not receive the intervention, i.e., a *control group*, as a powerful design element for guarding against threats to internal validity (Lee-St. John, 2012). The authors then lay out the concept of *random assignment*, the process by which adding a control group eliminates threats to internal validity (Kaplan, 2009; Lee St. John, 2012). In the context of experimentation, random assignment, or randomization, means that study elements have an equal probability of being placed into the treatment group or the control group (Kaplan, 2009). Given that *random assignment* is used to assign individuals to treatment conditions, selection, maturation, history, testing, and statistical regression are eliminated as plausible threats to internal validity over an infinite number of random assignments (Kaplan, 2009; Lee-St. John, 2012; Shadish, 2010).

***Rubin-Holland* causal model and treatment effect**

While Campbell greatly advanced our conceptual understanding of causal inference, it was Paul Holland (1986) and Donald Rubin's (1974) writings on causality that provided the statistical underpinnings for testing causal claims across the sciences (Lee St. John, 2012; Kaplan, 2009). Similar to Campbell (1957), their work is premised upon the counterfactual definition of causation first introduced by Hume and then later refined by the works of Mill, Mackie, and others (Lee. St. John, 2012; Kaplan, 2009; Hulswit, 2002; Peterson, 1898; West & Thoemmes, 2010). As such, Rubin and Holland also define the statement, "*X causes Y*" by its counterfactual, "*if X had not occurred, then Y will not have occurred*" (Kaplan, 2009; West & Thoemmes, 2010). The core concept of the Rubin-Holland model is that of *potential outcomes*, and ergo *the cause*, *X*, is conceptualized as being relative to another cause, including (but not limited to) the possibility of "*not X"* (Kaplan, 2009; Lee St. John, 2012; West & Thoemmes, 2010). The key distinction of this model from Campbell's is that the concept of *potential outcomes* is invoked to obtain precise estimates of the magnitude and direction of the causal effect via formal mathematical argument (West & Thoemmes, 2010).

The Rubin-Holland model originated from Neyman's nonparametric model concerning units with two potential outcomes, one observed and one unobserved (Glynn & Quinn, 2007; Kaplan, 2009). To begin, the Rubin-Holland model first considers a single unit (e.g., a human individual), denoted i, providing an outcome that is measured without error under at least two different treatment conditions within the same exact context (Lee-St. John, 2012; Kaplan, 2009; Morrison, 2011; Wet & Thoemmes, 2010). Rubin and Holland then define the causal effect as the difference between this individual's outcomes under the different treatment conditions, written as:

$$A(i) = Y_t(i) - Y_c(i) \qquad (1)$$

where $A(i)$ is the causal effect of interest, $Y_t(i)$ is the observed outcome for individual *i* under treatment *t*, and $Y_c(i)$ is the observed outcome for the same individual *i* under an alternative treatment condition, *c* (Lee-St. John, 2012; Kaplan, 2009; Hernan & Robins, 2018). Thus, we see that the Rubin-Holland model starts with the fundamental idea of causal inference at the individual level; however, Holland (1986) points out that this conception is patently flawed, as it is impossible to observe these two outcomes for the same unit, i. In other words, the condition of unit *i* receiving treatment precludes any possibility of unit *i* then being in the control condition within the same exact context, and vice versa (Kaplan, 2009; Lee-St. John, 2012). Holland refers to this problem as the *Fundamental Problem of Causal Inference* (Kaplan, 2009; Lee-St. John, 2012).

Rubin and Holland's solution to this problem was to move away from estimating a treatment effect at the individual-level and instead make inferences of the population of individuals (Lee-St. John, 2012; Kaplan, 2009; West & Thoemmes, 2010). In doing so, they shift their focus to summarizing treatment effects, i.e., $A(i)$'s, in the population, thus implicitly acknowledging that variability exists in treatment effects across *i* units. This framed causality as probabilistic, and consequently the *average treatment effect*, otherwise known as the *ATE*, became a statistic of central importance in causal inference methodology (Kaplan, 2009; Lee-St. John, 2012). The starting idea behind the *ATE* is that the aforementioned individual-level causal effect, $A(i)$, is a quixotic notion of causality, useful only in that it serves as a heuristic for thinking about research design. However, by leveraging this concept, Rubin and Holland arrive at useful approximations of this ideal, the most notable of which is the randomized experiment (Holland, 1986; Kaplan, 2009).

Beginning with two treatment groups, *A* and *B*, Rubin and Holland (1987) calculate the idealized group-level treatment effect as:

$$\mu_t(A) - \mu_c(A) \quad (2A)$$

$$\mu_t(B) - \mu_c(B) \quad (2B)$$

In this model, $\mu_t(A)$ and $\mu_t(B)$ represent the average outcomes for groups *A* and *B*, respectively, under treatment condition *T*. Equations 2A and 2B both represent average causal effects; however, it is important to note that these average causal effects may not be equal, as *A* and *B* may be representative of different populations (Hernan & Robins, 2018; West & Thoemmes, 2010). Furthermore, just like with the individual-level scenario, we can never observe the same group under both treatment conditions;

therefore, Rubin and Holland (1983) replace Equations 2A and 2B with the observed group difference:

$$ATE = \mu_t(A) - \mu_c(B) \quad (3)$$

For Equation 3 to serve as an accurate estimate for the group differences specified by Equations 2A and 2B - i.e., the group-level treatment effect – Rubin notes that additional assumptions must be made (Kaplan, 2009; Lee-St. John, 2012; West & Thoemmes, 2010). Identifying the necessary yet minimally sufficient set of assumptions for estimating an unbiased treatment effect via Equation 3 then becomes the central focus of Rubin and Holland's work (Kaplan, 2009; Lee-St. John, 2012). Of particular importance is the null effects selection assumption (NSEA), which states that in order for the ATE to be unbiased , i.e., $\mu_t(A) - \mu_c(B) = \mu_t(A) - \mu_c(A) = \mu_t(B) - \mu_c(B)$, the two treatment groups, *A* and *B*, must be identical to one another pre-treatment (Lee St. John, 2012; Rubin, 1983; West & Thoemmes, 2010); furthermore, we note that NSEA is a group-level assumption and thus the equivalence relation applies to group characteristics (Lee-St. John, 2012). In other words, two groups *A* and *B* are considered identical when they are similar in composition on both observed and unobserved covariates (Lee.-St. John, 2012; Kaplan, 2009). The main advantage of random assignment is that it guarantees that groups are matched on covariates over repeated randomization, thus ensuring NSEA (Deaton & Cartwright, 2017; Kaplan, 2009; Lee-St. John, 2012). As a result, randomization, in conjunction with other assumptions being met, ensures unbiased estimation of the average causal effect.

Given that only one of two potential outcomes can be observed per sampling element, we can conceptualize the counterfactual outcomes as a missing data problem.

Randomized experiments, like all other study designs, generate these missing data, allowing for only the observation of the outcome associated with the treatment condition actually received (Hernan & Robins, 2018; Kaplan, 2009; West & Thoemmes, 2010). However, the key difference with randomized experiments is that the random assignment mechanism embedded within such a design ensures that these missing data are *missing completely at random* (MCAR) (Hernan & Robins, 2018). This then allows estimation in spite of the missing data. More specifically, the assignment mechanism ensures that both observed and unobserved covariates are equally distributed across groups and everything but the treatment condition is the same (Kaplan, 2009; Lee-St. John, 2012). As a result, because the context to which the treatment is being applied is the same – i.e., selection is independent of covariates and the groups are identical – the treatment effect will manifest itself the same. In simpler terms, the treatment effect will be the same regardless of which group the treatment is actually administered to. This is the notion of *exchangeability*, and we more formally state this as $Y^a \perp\!\!\!\perp A$, $\forall$ a *in* A. When the treated and untreated are exchangeable, we say that the treatment is *exogenous* and can use the observed responses associated with receiving "not treatment" as the counterfactual for receiving treatment (Hernan & Robins, 2018; Kaplan, 2009). Subsequently, $\mu_t(A) - \mu_c(B) = \mu_t(A) - \mu_c(A) = \mu_t(B) - \mu_c(B)$ and the *ATE* provides us with an unbiased average treatment effect (Hernan & Robins, 2018; Shadish, 2010).

As outlined above, any serious attempt to address questions pertaining to causality require that the variation in treatment be exogenous (Gerring, 2011; Hernan & Robins, 2018; Pearl, 2009). Exogenous in the particular context of experimentation means that the assignment of students to treatment conditions is related to external causes (i.e., external to the model) and thus does not depend on the outcome variable or any variables related to the outcome (Stock & Watson, 2014). However, in the case of research situations where study elements are assigned to treatment and control through some non-random assignment process, the treatment assignment may no longer be independent of the outcome variable or its correlates. Moreover, important factors related to both treatment assignment and the outcome variable often cannot be captured by the types of statistical models used to estimate the average treatment effect. Consequently, any treatment assignment variation that is associated with these unmeasured confounders is subsumed into the residual, creating a correlation between the condition membership indicator and the error term. This correlation has been termed *endogeneity*. Moreover, this concept is important to consider because the linear statistical models commonly used in causal inference research carry with them the assumption that the regressors (e.g., the variable indicating membership in the treatment or control group) are exogenous. Otherwise, estimates for the parameter(s) of interest can become inaccurate, leading to faulty decision-making.

In summary, causal inferences in the absence of random assignment can be problematic and researchers continue to face challenges when aiming to evaluate the effects of a treatment or intervention when random assignment is not feasible. The following section describes how the instrumental variable regression approach can be used to address this issue, describes some of the weaknesses of the approach, and ends with a discussion of the methodological advancement that is proposed in this dissertation research.

Study Purpose: Extending the Instrumental Variable Approach

Instrumental variable methods are a promising solution for dealing with the aforementioned problem of endogeneity (Ebbes, 2004; Hueter, 2016). Augmenting the traditional linear regression outcomes model in which the outcome variable (Y_i) is regressed on the treatment variable (X_i) for study element *i* (e.g., $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$), this method introduces a variable, Z_i, that partitions out the problematic variation in the endogenous regressor (X_i). The process is conducted in two stages with a set of linear equations written as follows:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \varepsilon_i \quad \text{(second-stage model)} \tag{4}$$

$$X_i = \Pi_0 + \Pi_1 Z_i + \upsilon_i \quad \text{(first-stage model)} \tag{5}$$

Notably, the set of linear equations contains a model for the endogenous regressor, X_i, which is expressed as a linear function of Z_i plus error, υ_i. To be effective, the variable Z_i should (a) explain part of the variability in the endogenous regressor X_i; and (b) be uncorrelated with the error term in our regression outcomes model, ε_i. The latter assumption implies that the variable Z_i has no *direct* effect on the outcome of interest (Ebbes, 2004). If the variable Z_i satisfies these criteria, it is called an *instrument* (Ebbes, 2004; Hueter, 2016).

It is at this juncture that one can begin to see the crucial importance of Z_i and its associated properties. When Z_i is related to the endogenous regressor, X_i, its effect will be captured by the fixed, systematic component of the first-stage model in Equation 5, $\Pi_0 + \Pi_1 Z_i$. Of equal importance is the criterion that Z_i will be uncorrelated with the error term, ε_i, in the subsequent second-stage outcomes model in Equation 4. If this property holds, it then follows that the fixed, systematic component of the first-stage regression model, $\Pi_0 + \Pi_1 Z_i$, provides us with the exogenous piece of X_i while the first-stage residual, υ_i, represents the endogenous and problematic association between X_i and the outcomes model error term, ε_i.

Under this approach, the variability in X_i is decomposed into two pieces and the aforementioned problem of endogeneity can be addressed by using the predicted values of X_i, denoted $\hat{X}_i$, from the first-stage regression in Equation 5 as the new exogenous, explanatory variable for predicting Y_i in the second-stage outcomes model. It is important to note that in reality, the two sets of equations are modeled simultaneously and are only presented here in a stepwise fashion to aid in the explanation.

The challenge with this approach however, is that it assumes that a suitable instrumental variable is available to the researcher. In reality, such instruments can be difficult to obtain and no clear guidelines exist for how to identify a valid instrument (Ebbes, 2004). As a result, researchers often rely on their intuition and content knowledge to identify a useful instrument. Moreover, even when an instrument is identified, it often correlates poorly with the endogenous regressor, explaining only a small proportion of the variability in X_i, and therefore serves as a 'weak' instrument (Ebbes, 2004). Numerous studies have demonstrated that using weak instruments in IV analyses can result in potentially worse bias and inconsistency in the parameter estimate for the treatment effect than if the treatment estimate had simply been obtained from a traditional OLS regression analysis with no instrument at all (Ebbes, 2004; Hueter, 2016). In other words, the quality of the treatment effect estimate under the IV approach is highly dependent upon both the availability of the instrument as well as its quality.

In an effort to avoid this, researchers often identify observed instrumental variables that correlate strongly with the endogenous regressor X_i when aiming to identify a valid instrument. However, the very act of choosing an observed instrument that is highly correlated with the endogenous regressor X_i calls into question the assumption that the instrument itself is uncorrelated with the outcomes model error term (Ebbes, 2004). As such, endogeneity may still be present in the model. This issue makes the use of observed instruments extraordinarily difficult in applied research (Ebbes, 2004).

Promising Solution: Instrument Free Methods

A proposed solution to the issue faced by the use of observed instrumental variables are statistical approaches that require no observed instrument. Referred to as instrument-free methods, these approaches allow researchers to apply unobserved instruments in an IV approach that can support causal inferences. This dissertation research aimed to quantitatively evaluate an integrated student support model by exploring and extending the application of two statistical instrument-free methods for dealing with the endogeneity problem; these are the Latent Instrumental Variable (LIV) approach developed by Peter Ebbes (2004) and the Gaussian copula approach developed by Park and Gupta (2012).

Ebbes' LIV model adopts a mixture modeling approach to introduce an unobserved discrete binary variable that partitions the problematic endogenous regressor into two parts: an endogenous piece correlated with the structural error term, ε_i, in the outcomes model, and an exogenous piece (Ebbes, 2004). The Gaussian copula approach uses a copula function to model the joint distribution of the endogenous regressor and the structural equation error term (Park & Gupta, 2012). Under these two approaches, researchers can use the IV method without having an observed instrument, thus avoiding the issues of instrument availability, quality, and validity.

Despite the promise of the LIV and Gaussian copula approaches, an examination of the extant literature reveals that their empirical applications have been limited, and neither approach has been adopted for use in educational evaluation research. Moreover, comparisons among the Gaussian copula approach, the LIV approach, and the traditional IV approach are scarce, and a comparison using real-world randomized control trial data has never been conducted. In response to the dearth of research in this area, the goal of this dissertation was to both investigate the efficacy of the City Connects integrated student support model by utilizing the Gaussian copula and LIV approaches and to compare Gaussian copula and LIV regression to more classical approaches. The research questions are as follows:

1. How does estimation performance under the two-stage least squares IV (2SLS-IV) approach, the Latent Instrumental Variable (LIV) approach, and the least squares Gaussian copula approach compare across a range of research conditions involving endogeneity bias?
2. Using data from a real-world school lottery study examining the effect of the City Connects model of integrated student support, how do treatment effect estimates compare under the traditional 2SLS-IV approach with simulation-based propensity scores, the Latent Instrumental Variable (LIV) approach, and the least squares Gaussian copula approach? And, how do the model

parameters generated by instrument-free approaches compare to the observed instrument?

Data from two sources will be used to address the research questions. First, simulation data generated from different data generating processes will be used for exploring the performance of instrument-free methods and providing previously unexplored direct comparisons across a variety of different research conditions. Second, data collected as part of a large-scale study examining the effects of an integrated student support model on student outcomes will be used. This real-world evaluation using instrument-free methods will be valuable because a strong and valid instrument, a random offer from a lottery assignment process, is available to compare the three IV approaches. The application of the Gaussian copula and LIV methods with lottery data will also allow researchers to examine and triangulate the causal effects of an integrated student support on student academic achievement, providing important insights into key contributors of student success. Furthermore, the use of instrument-free methods will allow researchers to examine the causal effects of integrated student support with a much broader sample than allowed by traditional IV analysis, thus addressing possible self-selection bias that may arise due to systematic differences between lottery participants and non-participants and strengthening causal claims.

Summary

This chapter introduced the problem of causal inference from observational data followed by recently developed methods for causal inference that can be useful in settings where random assignment to experimental conditions is not possible. A review of the assumptions needed for justifying causal statements shows that ordinary regression techniques are ill-equipped for providing useful information about program impacts unless certain design conditions are met. Given that experimental designs are rare in program evaluation research, there is a need to evaluate the applicability of new statistical developments in the context of evaluation. The following chapter provides a theoretical overview of linear regression estimated via Ordinary Least Squares and Maximum Likelihood Estimation and Instrumental Variable regression along with limitations of these approaches. The researcher will then discuss two new causal inference methodologies for addressing the issues raised in both this chapter and the following.

CHAPTER 2: LITERATURE REVIEW

Linear regression was first developed and used to deduce cause-and-effect relationships by renowned German mathematician and physicist, Carl Friedrich Gauss (Freedman, 1997; Freedman, 2005). Proposed as "the method of least squares", Gauss used the method for estimating the orbital elements of astronomical objects (Freedman, 1997; Freedman, 2005). Subsequently, social scientists have used regression models for supporting causal statements for over a century now; however, Gauss' justification of the method of least squares, i.e., it is the unbiased linear estimate with minimum variance, is based upon conditions that are not easily achieved within social science contexts (Freedman, 1997). Measurements within the physical sciences, for instance, can often be made with great precision; furthermore, all relevant variables and the functional form of equations expressing orbital elements could be completely determined by Gauss via Newtonian mechanics (Freedman, 1997). Conversely, social scientists are far less certain about the relevant variables associated with an effect, and can rarely, if ever, measure all the variables relevant to a specified social-behavioral phenomenon. The result is that regression models applied to social science data are often either under-specified or misspecified, failing to control for important variables. Model specification problems are important to consider because the basic assumption behind regression methods in causal inference is that statistical control can replace random intervention, namely exogeneity (Gelman & Hill, 2006; Freedman, 1997). Thus, the enterprise of causal inference is inherently difficult within a social science context. To address these issues, more advanced regression-based techniques have been developed and applied.

The purpose of this chapter is to first introduce Ordinary Least Squares (OLS) linear regression as a foundation for the proposed methodological developments in this dissertation, outlining the method and its limitations in dealing with observational data. Instrumental Variable (IV) regression is subsequently presented as a common but often problematic alternative to OLS. This chapter ends with a discussion of the more recent instrument-free methodological developments for addressing endogeneity bias in causal inference research.

The Linear Model

The classic linear regression model that relates a continuous response variable, Y_i, and a discrete or continuous regressor, X_i, can be written as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \text{ for } i = 1,2,\ldots,n \qquad (6)$$

We first note that the response variable Y takes on different values, which is denoted by the subscript i. In the context of regression, this variation in the response variable, Y_i, can be partly explained by the variation in other variables, X (Van De Geer, 2005). Therefore we model Y_i as a linear function of X plus noise, which is denoted ε and reflects the variation in Y that is due to sampling error. The linear function of X we use for modeling Y is called the regression function and is comprised of observed covariables X and regression parameters, β, representing how the expected response in Y linearly depends on the covariables (Rodriguez, 2007). Given that the covariables are observed, the goal is to estimate the regression parameters from the sample data (Van De Geer, 2005). Thus we seek to find estimates of β for Equation 6, $\hat{\beta}$, that provides the best fit to the data, with "best fit" in this context being defined as the function that produces the least error, i.e., minimizes ε. Recalling that ε is the error resulting from a certain model specification, we can re-write it as $\varepsilon(\beta_0, \beta_1)$, representing it as a function of the parameter values. Thus, Equation 6 will be re-written as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon(\beta_0, \beta_1) \quad (7)$$

And solving for $\varepsilon(\beta_0, \beta_1)$ yields

$$\varepsilon(\beta_0, \beta_1) = Y - \beta_0 + \beta_1 X \quad (8)$$

Therefore, across *n* observations, we are interested in $\sum_{i=1}^{n} Y_i - \beta_0 + \beta_1 X_i$. However, due to errors being both positive and negative and canceling each other out in the summation, we square the deviation scores to arrive at the following function:

$$\sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 X_i)^2 \quad (9)$$

Parameter values for β are then chosen through the minimization of the function given by Equation 9, which is known as an objective function.

Maximum Likelihood Estimation (MLE)

Estimation of the regression parameters can also be achieved through the Maximum Likelihood Estimation (MLE) method, which is important to consider because it is a general-purpose estimation technique that allows for the linear model to be extended for different applications, including the use of instrument-free methods later discussed. A key difference under the MLE framework is that we assign a stochastic model to the error term of the regression model, whereas OLS requires no such stochastic assumption and instead is a distance-minimizing approach. Given that the continuous dependent variable, Y_i , will take on a range of different values, even for those sharing the same values on the covariates (i.e., being identical in characteristics), we assume that each observation, y_i, is a realization of a theoretical underlying random variable, Y_i, to which we assign a probability distribution (Rodriguez, 2007). For the classical linear model, the assumption is that the random variable has a normal distribution with mean, μ_i, and variance, σ^2, represented as follows:

$$Y_i \sim N(\mu_i, \sigma^2) \tag{10}$$

In Equation 10, μ_i represents the expected outcome for unit i, and σ^2 represents how much an actual observation may deviate from expectation (Rodriguez, 2007). It is important to emphasize here that the theoretical underlying random variable to which we assigned a probability distribution takes into account the fact that the response variable will take on a range of values for those who are identical on observed variables, and thus it reflects a conditional distribution. In other words, we are conceptualizing each observed response as one of many values we could observe under identical circumstances and therefore Equation 10 can be re-expressed as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \text{ where } \varepsilon \sim N(0, \sigma^2) \tag{11}$$

The normal distribution has a probability density function, written as follows:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu_i)^\wedge 2}{2\sigma^2}} \tag{12}$$

Equation 12 represents the distribution for one observation, y_i. However, through the assumption of mutually independent observations, the joint distribution of the data (i.e., all observations) is obtained by taking the continued product of the individual probability distributions given by Equation 12. This operation then leads to a likelihood function and allows for Maximum Likelihood Estimation (MLE) of the regression parameters β via standard calculus. Using the probability density function for a normal distribution, the likelihood function is written as follows:

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}} \tag{13}$$

An examination of Equation 13 reveals that the maximization of the likelihood function relies on minimizing the expression in the numerator above the exponential, and one should further note that this expression corresponds exactly to the objective function in OLS. As a result, the estimators given by the MLE and the OLS methods are equivalent under the assumption of normality.

Exogeneity Assumption of Regression and Bias

In causal inference, the regression parameters to be estimated, β, are of direct interest. When certain research design conditions are met, the parameter estimate associated with a given treatment regressor variable yields the causal effect of the intervention on an outcome of interest, *Y*. Thus, it becomes understandable that it would be useful for the parameter estimators to have certain desirable properties, such as accuracy and reliability (i.e., unbiasedness and consistency). However, properties of these estimators rely on crucial assumptions that come directly from the Gauss-Markov theorem, a proof proposed by mathematicians Carl Friedrich Gauss and Andrey Markov (Ebbes, 2004). The theorem states that the linear regression model has deterministic design matrix X and vector ε of uncorrelated errors with a mean of zero and the same finite variance, σ^2, as also stated by Equation 11 (Ebbes, 2004). If such conditions hold, the OLS estimator for β, which is given by the following equation:

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'Y, \quad (14)$$

is unbiased, consistent, and efficient (Ebbes, 2004; Hueter, 2016). If these assumptions are unmet, however, the inferential integrity of OLS estimators may be compromised.

A critical aforementioned assumption that must be revisited is that the vector ε has mean zero, i.e., $E(\varepsilon) = 0$. Such an assumption, along with that of a deterministic design matrix X, dictates that the explanatory variables in the model are uncorrelated with the error term, formally stated as $\rho_{\varepsilon,x} = 0$. It is imperative to mention that the condition $E(\varepsilon) = 0$ alone does not imply that $\rho_{\varepsilon,x} = 0$. Instead, $E(\varepsilon) = 0$ implies $\rho_{\varepsilon,x} = 0$ if and only if $E(\varepsilon X) = 0$. However, since the additional assumption of the regression model is that the design matrix is deterministic, the condition $E(\varepsilon X) = 0$ then holds, and thus $\rho_{\varepsilon,x} = 0$. This result is derived directly from the rules of expectation, where $E(aC)$ $= C(E(a))$ for some constant C and random variable a. It then follows from this result that the classic linear regression model assumes $\rho_{\varepsilon,x} = 0$. It is furthermore important to note that the assumption of the design matrix X being fixed across repeated samples (i.e., that it is deterministic) is often considered inappropriate for non-experimental science where researchers exercise much less control over predictor variables (Hueter, 2016). This problem is addressed by restating the assumptions of the Gauss-Markov theorem as being conditional on X (Hueter, 2016). Therefore, the assumption about the vector ε mean then moves from being $E(\varepsilon) = 0$ to $E(\varepsilon|X) = 0$. Yet, we note that the condition $E(\varepsilon|X) = 0$ implies that $E(\varepsilon X) = 0$, and thus the model still dictates that $\rho_{\varepsilon,x} = 0$, as the secondary and necessary condition for such a result still holds.

In sum, we see that properties of the OLS estimator were derived assuming X is fixed and thus cannot correlate with the error, which is reasonable to assume with data obtained from experiments. However, in moving toward a framework with random regressors (i.e., observational studies where researchers play a much more passive role), both the error and the predictor variables are now considered random variables and can correlate; therefore an assumption is placed on the relationship between the regressors and the error term to provide results for the OLS estimator that are identical under the assumption of a deterministic design matrix. However, if $\rho_{\varepsilon,x} \neq 0$, then X is said to be *endogenous*, and the OLS estimator for β may then be biased, inconsistent, and inefficient, losing all properties that made it desirable as an estimate of the population parameter (Ebbes, 2004; Hueter, 2016).

A major problem resulting from endogenous regressors is that of inconsistency. In mathematical terms, consistency is represented as:

$$\operatorname*{plim}_{n \to inf} \hat{\beta}_n = \beta \tag{15}$$

Equation 15 implies that as the sample size approaches infinity, the variance of the estimator, a random variable, goes to zero and the estimates converge on the true parameter value (Ebbes, 2004). In practical terms, consistency means that the estimate becomes more accurate with more data. In causal inference this is important, as inconsistent regression estimates measure only the magnitude of an association, failing to capture both the magnitude and direction of causation (Stock & Watson, 2014). Returning to the standard linear regression model, the problem with endogenous regressors can clearly be illustrated. The standard regression model specifies a continuous response (y), regressors (x), error term, (ε), and OLS estimate $\hat{\beta}$ of β, reflecting the deviation from the conditional mean for every exogenous change in the regressor, *x,* provided $\rho_{\varepsilon,x} = 0$ (Ebbes, 2004). If x is uncorrelated with the error term ε, then the only effect x has on y is a direct effect, as seen below:

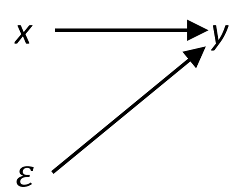

Figure 1. Exogenous Regressor

However, if there is an association between the regressor and error term, i.e., $\rho_{\varepsilon,x} \neq 0$, then *x* has both a direct *and* indirect effect on *y*, shown as follows:

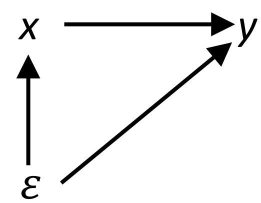

Figure 2. Endogenous Regressor

In this case, there is both a direct effect, βx, and an indirect effect, ε, which through its effect on *x*, affects *y.* As a result, changes in *x* now have two effects on y, and the OLS estimate yields either $\hat{\beta} > \beta$ or $\hat{\beta} < \beta$. Standard calculus reveals that the total derivative of $y = \beta x + \varepsilon$ with respect to x, yields the following:

$$\frac{dy}{dx} = \beta \tag{16}$$

However, when endogeneity bias is present, any change in *x* is also a function of the error, and thus the total derivative of the following is taken:

$$y = \beta x + \varepsilon(x) \tag{17}$$

Standard calculus now yields

$$\frac{dy}{dx} = \beta + \frac{d\varepsilon}{dx}, \tag{18}$$

where $\frac{dy}{dx}$ is estimated from sample data and the OLS estimate is now the net effect $\beta + \frac{d\varepsilon}{dx}$. Under these circumstances, the estimate is likely to be both unreliable and inaccurate.

The Two-Stage Least Squares Instrumental Variable (2SLS-IV) Method

A solution to the endogeneity problem is to use an experimental design where assignment to the conditions on *X* is based on a random process. However, it is often the case across research settings that randomization is either unfeasible or even unethical (Adelson, 2013; Ebbes, 2004). A proposed solution has been the instrumental variable regression approach, spearheaded by economists as a panacea to the endogeneity problem that has plagued nonexperimental science. Provided one has an instrument, *Z*, that is both correlated with the endogenous regressor, *X*, but uncorrelated with the regression model error term, ε, one can model a simultaneous set of linear equations, $Y = X\beta + \varepsilon$ and $X = Z\Pi + \upsilon$, to circumvent the problem of endogeneity and arrive at consistent estimators (Ebbes, 2004; Hueter, 2016). The core idea is that *X* is partitioned into an exogenous random variable, $Z\Pi$, and an endogenous random variable, υ (Ebbes, 2004).

Assuming a set of instrumental variables are available, regression parameters for the IV model can be estimated by means of simultaneous equation estimation techniques. The most widely used estimation techniques are Limited Maximum Likelihood Estimation (LIML) and Two Stage Least Squares (2SLS), of which only the 2SLS estimator will be discussed and presented here due to its simplicity and ubiquity. Given the observed instrument, *Z*, the unobserved Π and $Z\Pi$ in the previously presented set of linear equations are estimated by first regressing the endogenous explanatory variable X on the observed instrument, Z (Ebbes, 2004; Hueter, 2016). In the context of IV regression, Z is exogenous and thus its covariation with X comprises exogenous variation, namely $Z\Pi$. We recall from Equation 14 that ordinary least squares regression of a variable Y on a given design matrix X produces the following OLS estimator:

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'Y \qquad (19)$$

The first-stage regression of the endogenous regressor X on Z is also an ordinary least squares regression, and thus the OLS estimator for this regression is

$$\hat{\beta} = (Z'Z)^{-1}Z'X \tag{20}$$

We note the structural similarities between Equations 19 and 20 and also note from Equation 20 that $(Z'Z)^{-1}Z'$ is just the usual OLS regression projection matrix, denoted P_z, mapping the first-stage response variable, *X,* to the predicted values produced by the linear function, $\hat{X}$ (Hueter, 2016). Information contained by this projection matrix is then encoded into the second-stage linear regression, modifying the OLS estimator to give the following IV estimator:

$$\hat{\beta}^{IV}=(X'P_zX)^{-1}X'P_zY \tag{21}$$

In sum, the 2SLS-IV approach is equivalent to regressing the endogenous regressor on the instrument, saving the predicted values, and subsequently using these predicted values in the second-stage outcomes model to produce a regression coefficient detailing the linear relationship between these predicted values of *X* and the dependent variable, *Y* (Ebbes, 2004; Hueter, 2016).

Instrumental Variables in Education Research: Lottery Studies

Recently, the econometric literature has begun to examine the effect of school choice (e.g., pilot schools) on student academic outcomes via lottery studies. Essentially, this has been achieved by capitalizing on the random component embedded within many state's school assignment mechanisms. A naïve evaluation of school effects is problematic due to selection bias, as many non-random factors determine where students attend school, and thus endogeneity bias is extremely likely (Abdulkadiroglu, Angirst, Dynarski, Kane, & Pathak, 2011). However, a number of major cities, such as Boston and New York City, have recently begun to assign students to schools using a centralized assignment process based on the Gale-Shapley Deferred Acceptance algorithm

(Abdulkadiroglu, Angirst, Narita, & Pathak, 2017). Within this process, if the number of applicants to a school is larger than the number of available seats at that school, then a random process is invoked. More specifically, students rank their school choices in terms of preference, with each student only being allowed a set maximum number of choices. Applying students are then assigned a priority category at each school, which can be looked at as the schools' preferences over students. For students selecting the same school (i.e., tied on school preference) and belonging to the same priority group (i.e., tied on ranking within that school), a randomly generated lottery number is used to break the tie and determine who gets placement. Thus, conditional on application cycle, school preference, and priority group, which has been referred to as a "risk set" in the lottery literature, this system produces random assignment with known probabilities (Abdulkadiroglu et al., 2017). This conditional random assignment can then be capitalized on to create an instrument indicating whether or not a student received a random offer to attend one of their preferred schools. Subsequently, the random offer can be used as an instrumental variable in a 2SLS-IV linear regression model as follows:

$$Y_i = \alpha_2 + \sum_j \delta_j d_{ij} + X_i' \beta + \beta_{School} \widehat{School}_i + \varepsilon_i \tag{22A}$$

$$School_i = \alpha_1 + \sum_j \kappa_j d_{ij} + X_i' \Pi + \Pi_{IV} Z_i + \eta_i, \tag{22B}$$

Where Y is some student outcome of interest, $\sum_j \delta_j d_{ij}$ are the risk sets previously introduced and used to create conditional random assignment, $School$ is the variable indicating attendance at a particular school, and Z is a dummy coded variable indicating whether or not a student received a random offer to attend his/her school of choice. In the presence of selection bias, the school attendance variable is endogenous and therefore Z is used to partition school attendance into exogenous and endogenous variation.

Limitations of Instrumental Variable Approaches

While promising in theory, the IV regression method has notable shortcomings in practice. Most notably, the method assumes an appropriate instrument, which can be difficult for several reasons (Crown, Henk, & Vanness, 2011). First, no clear guidelines for how to find valid instruments exist, leaving researchers to rely on intuition and theory and essentially "guess" at what may serve as a good instrument (Ebbes, 2004). Secondly, the available instrument must satisfy demanding criteria, which relate to relevance and exogeneity (Ebbes, 2004; Hueter, 2016; Stock & Watson, 2014). Instrument relevance dictates that the instrumental variable is strongly correlated with the endogenous regressor but uncorrelated with the outcome variable (Ebbes, 2004). In addition, the instrument must be exogenous and so uncorrelated with the error term; yet many researchers have noted the unlikelihood of both simultaneously occurring (Crown et al., 2011; Ebbes, 2004; Hueter, 2016). Crown et al. (2011), among others, argues that the stronger the association between the endogenous regressor and the instrument (i.e., the greater the instrument relevance), the more likely it is that the instrument is correlated with the error term; thus, in an effort to minimize this correlation with the error term, researchers have a tendency to identify weak instruments. Ebbes (2004) agrees that the features that make instruments exogenous are also likely to make instruments weak.

Bound, Jaeger and Baker (1995) prove the relative consistency of IV to OLS to be as follows:

$$\frac{\rho_{z,\varepsilon}/\rho_{x,\varepsilon}}{\rho_{x,z}} \tag{23}$$

A close examination of the denominator, $\rho_{x,z}$, reveals that when an instrument is weak, even a small amount of endogeneity bias can result in inconsistency that is potentially larger than that resulting from an OLS model (Bound et al., 1995; Ebbes, 2004). Even in the presence of exogeneity, the literature is replete with warnings about the use of weak instruments and the potential for the estimates to be inconsistent, estimated with far less precision (i.e., estimates with inflated standard errors), and more biased than those for OLS; in effect, they lose all properties that made them an attractive option in the first place (Hueter, 2016). For example, Bound, Jaeger and Baker (1995) revisited the results from a well-known study conducted by Angrist and Krueger (1991) where quarter of birth (i.e., season) was used as an instrument in examining the causal effect of educational attainment on wages in a large U.S. Census sample. The authors found evidence suggesting that the weak correlation between the instrument and the endogenous regressor was problematic enough to statistically significantly affect their estimates, and a significant finite sample bias for the reported estimates also existed, bringing into question the validity of the study's findings (Bound et al., 1995). This is but one example of the difficulty of finding a valid instrument, and, as a result, the consequences of a weak instrument.

Even in school choice studies where random lottery offers may arguably satisfy the relevance and exogeneity criteria, the disadvantages to relying on centralized assignment mechanisms to obtain an instrument are still numerous. First, obtaining valid lottery instruments from centralized assignment mechanisms requires researchers to develop expertise in the school assignment algorithm, which are district specific and can be fairly complex. If the assignment algorithm is unable to be replicated with a high level of accuracy, the instrument can be rendered inaccurate and weak, leading to the aforementioned problems of inconsistency and bias that threaten the internal validity of instrumental variable methods in general. Even with an understanding of the assignment process, the iterative nature of the school assignment algorithms make it hard to disentangle which students were randomized to schools. Abdulkadiroglu et al. (2017) note that traditional IV regression studies relying on lottery offers embedded within centralized assignment mechanisms have failed to capture the full random variation that exists within the assignment mechanism. Consequently, this weakens lottery instruments. Thus we see that while lottery studies can provide researchers with valid and useful instruments, instrument availability and quality largely depend on the researcher's ability to understand and replicate district-specific, complex assignment algorithms and subsequently capture the random assignment existing within these processes.

In situations where the instrument is both relevant and exogenous, the IV estimator is consistent; however, a fact often ignored by many applied researchers is that the IV estimator is biased in finite samples (Ebbes, 2004); moreover, in small samples this bias can be rather substantial. Incidentally, randomization in many school assignment mechanisms only takes place for a portion of students in the school assignment process, and, consequently, result in reduced sample sizes (Abdulkadiroglu et al., 2017; Steele, Slater, Zamarro, Miller, Li, Burkhauser, & Bacon, 2017). The primary reason for this is that the allocation system based on the student-proposing deferred acceptance algorithm employs a multi-stage market design, randomly assigning only the subset of students who are tied on both school preference and priority ranking (Abdulkadiroglu et al., 2017; Abdulkadiroglu et al., 2011). Traditionally, the majority of lottery study instruments have been observed by considering only oversubscribed schools and students sharing high-dimensional balancing scores known as *marginal priority groups*, which are discrete sets containing students with the exact same school preference list and priority rankings (Abdulkadiroglu et al., 2011; Abdulkadiroglu et al., 2017; Steele et al., 2017). Abdulkadiroglu et al. (2017) note that lottery study strategies involving identifying random offers via full stratification on school preferences and student priorities considerably reduces degrees of freedom and sample size, eliminating schools and students from the analytic sample.

The subsetting of school lottery data based on random offer instruments poses a potential threat to IV analyses, as smaller data sets induce greater variance and can substantially bias the estimates (Boef, Dekkers, Vandenbroucke, & le Cessie, 2014). Moreover, not all students partake in the school assignment processes underlying school lottery IV methods - a large group of students in some districts essentially opt out of the assignment process completely and get administratively assigned to a school. These students must then be excluded from the sample. Therefore a valid concern that may arise with the use of any observed lottery instrument is the possibility of systematic differences existing between the group of students/families that opt out of the school assignment process and those who choose to partake. In cities such as Boston, the school choice mechanism has had a history of associated problems, such as parents exhibiting strategic behaviors for gaming the system and implementation issues (e.g., walk-open precedence), which have led to unintended results and some general distrust of the system (Abdulkadiroglu, Pathak, Roth, & Sonmez, 2006). Additionally, school choice mechanisms can be daunting, presenting as a complicated web of options to many parents. As a result, it is reasonable to assume that some families may opt out of the school assignment process for non-random reasons, biasing the results we obtain from instrumental variable methods applied to data using only lottery participants.

Instrument-Free Methods

The proposed solution to the identified problems is statistical approaches that require no observed instrument, called instrument-free methods. Instrument-free methods allow researchers to address issues of endogeneity and thus make causal inferences without needing to identify and justify instruments, one of the key problems associated with the traditional instrumental variable method. Although a few instrument-free methods exist, the researcher specifically focuses on two recent and particularly promising instrument-free approaches, both of which have never been applied within the context of educational research nor have received mention in the educational research literature.

Latent Instrumental Variable Approach: A Promising Solution

The Latent Instrumental Variable (LIV) approach developed by Ebbes (2004) is similar to the classical IV approach in that it assumes the endogenous regressor can be partitioned into two pieces, an endogenous part and an exogenous part (Ebbes, 2004; Hueter, 2016). Unlike the classical IV approach, however, it does not rely on an observed instrument and thus circumvents all previously introduced issues of instrument availability and validity. The LIV model simultaneously estimates a dichotomous grouping of the data along with other parameters using mixture modeling techniques, treating the discrete latent instrument as a nuisance parameter to be integrated out across a finite mixture (Ebbes, 2004). Put simply, mixture models are a combination of two or more probability density functions or data generating mechanisms, which together then form a probability density function, known as a mixture. This model is a sum of individual probability density functions weighted by their respective mixing proportions, which are component priors that collectively sum to one (Shalizi, 2015). Given that each component function used in the summation is a probability density function, the resulting function is also a probability function through the property of convexity. The property of convexity refers to a linear combination of inputs where all inputs are weighted by coefficients that are non-negative and sum to one. One can view convex combinations as being subsumed under linear combinations, with the former preserving certain properties that make the resulting combination - i.e., mixture density - a probability density function as well. This is written as:

$$f(y_t) = \sum_{k=1}^{K} \pi_k f_k(y_t) \quad (24)$$

Given the above formulation, any N-dimensional continuous random variable, Y, is interpreted as being generated from K distinct random processes, respectively modeled by $f_k(y_t)$, each with π_k proportion of observations (Shalizi, 2015).

The structural form of the LIV model, as presented by Ebbes (2004), is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (25A)$$

$$x_i = \pi' \tilde{z}_i + v_i \quad (25B)$$

in which the unobserved instrument, $\tilde{z}$, is treated as being discrete (if it had been observed, it would separate the sample into *m* groups) and π is an (*m* x 1) vector of category means. For model identification purposes, the number of category means must be equal to or greater than two and must be distinct. Simulation results from a study conducted by Ebbes (2004) have shown that the simple LIV model (where $m = 2$) is robust against misspecification of the true number of categories and performs well overall. In the model, it is assumed that $\tilde{z}$ is independent of the error terms (ε, v). Furthermore, in Ebbes' formulation of the model, the joint error terms are specified to follow a joint normal distribution, denoted *H*, with a mean of zero and a variance-covariance matrix as follows:

$$\Sigma = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_v^2 \end{bmatrix} \tag{26}$$

where the correlation between the endogenous regressor and the outcomes model error term is captured by $\sigma_{\varepsilon v}$.

It is also assumed that the categories are unknown a priori and follow a multinomial distribution with parameters (n, λ), where $n = 1$. Furthermore, we rely on the simple LIV model. Given *n* i.i.d. observations (y_i, x_i), the marginal probability density function for (y_i, x_i) is given as

$$f(y_i, x_i) = \lambda f_1(y_i, x_i) + (1 - \lambda) f_2(y_i, x_i) \tag{27}$$

where f_j is the conditional normal bivariate probability density function, given $\tilde{z}_i$ = e_j (i.e., density for a subpopulation), and where e_1=$(1,0)'$and e_2=$(0,1)'$. Referring back to the previous discussion of mixture models and Equation 24, we note that $f(y_i, x_i)$ is a mixture of bivariate homoscedastic normal distributions; furthermore, the specification of the multinomial distribution for the latent instrument dictates that the parameters used as coefficients in Equation 27 are non-negative and sum to one (as they are proportions, i.e., fractions), ensuring the property of convexity holds and the resulting mixture is a density as well. It is important to emphasize that Equation 27 is just an averaging of conditional densities over all values of $\tilde{z}_i$, and thus we marginalize out the instrument, $\tilde{z}_i$, to arrive at an unconditional density for (y_i, x_i), removing the need for an observed instrument. This is a standard result from elementary probability, where one can use the Law of Total Probability to move from a weighted sum of conditional densities to a marginal density. Ebbes (2004) notes that this mixture distribution (Equation 27) has the following expectation

$$\mu_{y,x} = \begin{pmatrix} \beta_0 + \beta_1(\lambda\pi_1 + (1-\lambda)\pi_2) \\ \pi_1 + (1-\lambda)\pi_2 \end{pmatrix} \tag{28}$$

and variance-covariance matrix

$$\Omega_{y,x} = \Omega + \lambda(1-\lambda)\ (\pi_1 - \pi_2)^2(\beta_1, 1)'(\beta_1, 1), \tag{29}$$

where Ω is the following reduced form variance-covariance matrix

$$\Omega = \begin{bmatrix} \beta_1^2\sigma_v^2 + 2\beta_1\sigma_{\varepsilon v} + \sigma_\varepsilon^2(1-\lambda)\pi_2 & \beta_1\sigma_v^2 + \sigma_{\varepsilon v} \\ \beta_1\sigma_v^2 + \sigma_{\varepsilon v}\pi_1 & \sigma_v^2 \end{bmatrix} \tag{30}$$

For estimation of the parameters, we obtain the continued product of Equation 27 across all observations to arrive at the likelihood function (Ebbes, 2004). Subsequently, the method of maximum likelihood estimation can be used to estimate the model parameters, β_0, β_1, π_1, π_2, Σ, and λ. Note again that $\tilde{z}_i$ is left out, as the model does not require an observed instrument.

Gaussian Copula Method: A Promising Second Solution

Park and Gupta (2012) propose a model that identifies parameters through maximizing the likelihood resulting from the joint distribution of the endogenous regressor and structural equation error term. In order to arrive at the joint distribution of the endogenous regressor and structural error term and thus account for the correlation between the two quantities, Park and Gupta rely on a copula. A copula is a function linking a multivariate distribution to marginal univariate distributions (Papies et al., 2017; Park & Gupta, 2012). If we think of a *n*-dimensional unit cube $[0,1]^n$, then the copula is the multivariate distribution function defined on this space, with the marginal distributions for each variable being uniformly distributed. In mathematical terms, if we let $X = (X_{1,\ldots,}X_n)$ be a random vector with cumulative distribution function *F*, defined as $F = Pr[X_1 < x_1, \ldots, X_n < x_n]$, and with uniform marginal distribution functions F_i, such that $X_i \sim F_i$, then the distribution function *C* is called a copula of *X* if:

$$F = C(F_1, \ldots, F_n) \tag{31}$$

Thus we see that *C* is a dependence structure, and the multivariate distribution has been decomposed into two components: the copula function and the uniform marginal distributions (Embrechts, Lindskog, & McNeil, 2001; Park & Gupta, 2012). One of the most important results regarding copulas is Sklar's Theorem, which states that if the marginals, F_i, are all continuous, then the *n*-dimensional copula function *C* is uniquely determined (Embrechts, Lindskog, & McNeil, 2001). From a modeling standpoint, this then allows us to separate the univariate marginal distributions and the copula, capturing the dependence structure between the marginals with the copula, *C* (Embrechts, Lindskog, & McNeil, 2001).

The essential idea behind Park and Gupta's instrument-free method is to use information from the joint distribution of the endogenous regressor and structural error via a copula model to arrive at consistent estimators, thus obviating the need for an observed instrument (Park & Gupta, 2012). They achieve this by first selecting marginal distributions for the endogenous regressor and structural error term, which are based on information contained in the data and modeling assumptions, respectively (Park & Gupta, 2012). Subsequently, the copula model estimates a joint distribution of the error and endogenous regressor from the marginal distributions, allowing for a range of correlations between the two marginals in the process.

For the marginal density of the error term, denoted $g(\varepsilon)$, Park and Gupta (2012) specify a normal distribution, an assumption common to many linear modeling approaches. Unlike the structural error term, however, Park & Gupta note that data for the endogenous variable is observed (2012). As a result, they consider these observed data to be sample data from the true distribution of the endogenous regressor, *X*, and use a nonparametric density estimation approach to allow for the data to determine the marginal density function of the endogenous regressor (Park & Gupta, 2012). The proposed formula for the marginal density estimator is written as:

$$\hat{h}(x) = \frac{1}{Txb}\sum_{t=1}^{T} K(\frac{x-X_T}{b}) \quad (32)$$

Where we assume $X_1, \ldots\ldots, X_T$ to be i.i.d observations with true density *h(x)*, *b* is a data-driven bandwidth, and *K(m)* = 0.75 *x* (1-m^2) *x* *I*($|m| \leq 1$), where *I(y)* is an indicator function. Given specifications for the marginal densities, Park and Gupta then construct a joint distribution function from marginal distributions of the endogenous regressor and structural error, respectively denoted $H(x)$ and $G(\varepsilon)$. If we allow $F(x, \varepsilon)$ to be the joint distribution function of the endogenous regressor and error term, with marginal distributions as previously stated, then Sklar's Theorem states the following:

$$F(x, \varepsilon) = C(H(x), G(\varepsilon)) = C(U_x, U_\varepsilon) \quad (33)$$

Given that $H(x)$ and $G(\varepsilon)$ are marginal distribution functions, the probability integral transformations $U_x = H(x)$ and $U_\varepsilon = G(\varepsilon)$ in Equation 33 are $Uniform(0,1)$ random variables and *C* is a copula. To show this, let an arbitrarily chosen random variable *Z* follow a uniform distribution, denoted $Z \sim U(0,1)$, with the resulting probability density function specified as

$$f(z) = 1, 0 < z < 1. \quad (34)$$

Defining the Cumulative Distribution Function (CDF), denoted $F_Z(z)$, as $P(Z \leq z)$, we then have that

$$P(Z \leq z) = \int_0^z 1\, dz = z \tag{35}$$

Therefore $F_Z(z) = z$ for a $Uniform(0,1)$ random variable. Now let the endogenous regressor, a random variable *X*, have the continuous cumulative distribution function, *F,* and let $U_X = F(X)$. Then we have the following:

$$F_{U_x}(u_x), = P(U_x \leq u_x) = P(F(X) \leq u_x) = P(X \leq F^{-1}(u_x)) = F(F^{-1}(u_x)) = u_x$$

(36)

Therefore, U_x is a uniform random variable. The same arguments are used for $G(\epsilon)$**.** Given this result, we then see that the copula is a bivariate distribution function on the unit cube $[0,1]^2$.

The proposed copula model relies on a bivariate normal assumption of the variables to use a Gaussian copula, defined as:

$$C(U_x, U_\varepsilon) = \Psi_\rho(\Phi^{-1}(U_x), \Phi^{-1}(U_\varepsilon))$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(U_x)} \int_{-\infty}^{\Phi^{-1}(U_\varepsilon)} \exp[\frac{-(x^2-2\rho xy+y^2)}{2(1-\rho^2)}]\, dxdy, \tag{37}$$

where Ψ_ρ is the bivariate normal distribution function with parameter correlation coefficient ρ, and Φ is the univariate standard normal distribution (Park & Gupta, 2012). Differentiating Equation 37 yields the joint probability density function of the endogenous regressor and the structural error, represented as

$$f(x, \varepsilon) = \frac{\partial^2 C(U_x, U_\varepsilon)}{(\partial x \partial \varepsilon)}\, h(x) g(\varepsilon), \tag{38}$$

where $h(x)$ and $g(\varepsilon)$ are marginal densities for the endogenous regressor and structural error term, respectively, and have been discussed. Using this joint density function, one is then able to obtain a likelihood function, allowing for parameter estimation via Maximum Likelihood Estimation (MLE). However, given the aforementioned specifications, there is a much simpler way to estimate the model. The Gaussian copula assumes that the joint distribution follows a bivariate normal distribution (Park & Gupta, 2012). Therefore, letting $\Phi^{-1}(U_x) = X^*$ and $\Phi^{-1}(U_\epsilon) = \varepsilon^*$, we see that the proposed copula method dictates that X^* and ε^* are jointly distributed as a standard bivariate normal with correlation ρ (Park & Gupta, 2012). As a result, we can rewrite the model as:

$$\begin{pmatrix} X^* \\ \varepsilon^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad (39)$$

Where z_1 and z_2 are independent random variables drawn from a standard normal distribution (Park & Gupta, 2012). Application of standard matrix algebra yields:

$$X^* = z_1 \quad (40A)$$

$$\varepsilon^* = \rho z_1 + \sqrt{1-\rho^2} z_2 \Rightarrow \varepsilon^* = \rho X^* + \sqrt{1-\rho^2} z_2 \text{ (by definition of } X^*) \quad (40B)$$

Subsequently, if we recall the normality assumption of ε, and also recall that $\varepsilon^* = \Phi^{-1}(G(\varepsilon))$, which is just a linear transformation of ε, we then get:

$$\varepsilon = \sigma_\varepsilon \varepsilon^* + \mu_\varepsilon \Rightarrow \varepsilon = \sigma_\varepsilon \varepsilon^* \text{ (by assumption of error)}$$

$$\Rightarrow \varepsilon = \sigma_\varepsilon \rho X^* + \sigma_\varepsilon \sqrt{1-\rho^2} z_2 \text{ (by equality given in Equation 40B)} \quad (41)$$

We can now write the linear regression model with endogenous regressor *X* as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \Rightarrow Y_i = \beta_0 + \beta_1 X_i + \sigma_\varepsilon \rho X_i^* + \sigma_\varepsilon \sqrt{1-\rho^2} z_{2,i} \quad (42)$$

By including X_i^* in our linear regression model, we thus decompose the structural error into two components: $\sigma_\varepsilon \rho X_i^*$, the portion of the structural error correlated with the endogenous regressor, and $\sigma_\varepsilon\sqrt{1-\rho^2}z_{2,i}$, which is exogenous (Park & Gupta, 2012). We view X_i^* as the copula Control Function (CF), which controls for the dependence between the endogenous regressor and the structural error and, as a result, allows us to consistently estimate β using OLS (Papies et al., 2017; Park & Gupta, 2012). This approach is similar to the CF approach used with traditional instrumental variable analyses, where the residuals from the regression of the endogenous variable on the observed instrument are then input into the outcomes model as an additional regressor, thus controlling for the correlation between the two quantities (Papies et al., 2017). The least squares Gaussian copula model for dealing with endogeneity bias is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^* + \varepsilon_i, \tag{43}$$

where $X^* = \Phi^{-1}(H(X))$, and $H(X)$ is an empirical cumulative density function (CDF) estimated from the data using a rank ordering of the endogenous variable X and calculating the proportion of observed values that are less than or equal to each rank ordered value (Papies et al., 2017; Park & Gupta, 2012). An important advantage of the copula method over other IV approaches, including LIV, is that it imposes no exogeneity requirement on the instrument (Park & Gupta, 2012). However, an important modeling assumption under this approach is that the endogenous regressor is non-normally distributed, which should be empirically tested prior to estimation (Papies et al., 2017; Park & Gupta, 2012).

Conclusion

As this discussion shows, instrument-free methods are a useful way for addressing endogeneity bias from observational studies. Moreover, they are less restrictive than traditional solutions relying on observed instruments, which can be hard to identify and potentially result in model bias that is worse than if estimates had been obtained by means of OLS. In the chapter that follows, the proposed methods will be discussed along with specific research plans for addressing the dissertation research questions.

CHAPTER 3: RESEARCH DESIGNS AND METHODS

This dissertation research explored various methods for estimating the causal impact of City Connects and also examined the performance of these methods under a range of simulated research conditions involving endogeneity bias. Compared to the traditional 2SLS-IV approach, the LIV and the Gaussian copula instrument-free approaches are novel and underused approaches for dealing with endogeneity bias, and to-date, neither approach has been adopted for educational evaluation research. Furthermore, the adequacy of these methods compared to the traditional 2SLS-IV approach has not been fully explored nor demonstrated (Ebbes, 2004). At present, only one direct comparison of the LIV and Gaussian copula approaches can be found in the extant research literature, and this comparison examines the performance of the two methods under only a single endogeneity condition. Furthermore, comparing the estimation performance of instrument-free approaches and the classical IV approach will not only help researchers better understand the proposed instrument-free methods and the conditions under which they are most effective, but may also help to promote their use amongst applied researchers trained within the classical IV framework. Therefore, any opportunity to examine and compare the performance of these approaches with data involving endogeneity bias should be pursued, especially using real-world data arising from an RCT with a strong and arguably valid instrument, such as a school lottery assignment study. To the author's knowledge, this is the first ever comparison of instrument-free methods with IV under an RCT design in the field of education. Moreover, this research provides the first-ever evidence from an RCT demonstrating the efficacy of the City Connects intervention. The research questions that guided this research are as follows:

1. How does estimation performance under the two-stage least squares IV (2SLS-IV) approach, the Latent Instrumental Variable (LIV) approach, and the least squares Gaussian copula approach compare across a range of research conditions involving endogeneity bias?
2. Using data from a real-world school lottery study examining the effect of the City Connects model of integrated student support program, how do treatment effect estimates compare under the traditional 2SLS-IV approach with simulation-based propensity scores, the Latent Instrumental Variable (LIV) approach, and the Gaussian copula approach? And, how do the model parameters generated by instrument-free approaches compare to the observed instrument?

This chapter begins with a section describing the data sources used to address these research questions. To provide the necessary context for understanding the application of the instrument-free methods to the program evaluation of an integrated student support model, this section will also include a description of the intervention and the lottery assignment process. Lastly, methods and data analyses for addressing the research questions will be discussed.

Data Sources

This dissertation research used both synthetic and real-world program evaluation data to explore the performance of instrument-free methods for estimating the true causal effect across a range of endogeneity conditions. The synthetic data was artificially generated by the researcher and specified to represent certain conditions, while the program evaluation data comprises student lottery records, student demographics and student outcomes data, e.g., student GPA, all of which have been obtained directly from a large urban school district, henceforth referred to as District Z.

Synthetic data: Monte Carlo experiments

Simulated data was generated in the R Statistical Computing Environment under a set of data generating processes representative of various regression scenarios. The range of data generating processes were designed to highlight both the performance and boundaries of instrument-free methods compared with that of OLS and the traditional two-stage least squares IV approach (2SLS-IV). To achieve this goal, the method of inverse transform sampling was used to generate data according to six studies, all of which fall under the scope of Research Question 1 and will be outlined in further detail in the methods and data analysis section to follow.

School lottery data: Program Evaluation of City Connects Intervention

The program evaluation portion of this research will take advantage of a 'natural experiment' that occurred within District Z, where students were assigned to schools via a centralized lottery-based assignment mechanism based on a student proposing deferred acceptance algorithm involving a random component. The deferred acceptance algorithm is described in the Lottery Design section. Using real-world student and school lottery data, the causal effect of attending a school within District Z that received an integrated student support intervention, the City Connects intervention, on student academic achievement will be compared under the instrument-free methods and the traditional 2SLS-IV method. Since District Z uses a lottery mechanism in their school assignment process, the comparison among IV approaches will be conducted with a strong and valid observed instrument. Furthermore, the triangulation of results from empirically sound methods, one of which makes use of conditional randomization, makes a distinct contribution to the field of program impact evaluation, as this work draws on a combination of novel methods to offer insights into the causal impacts of an integrated student support intervention implemented within the District Z school system, namely City Connects. The causal effect of the City Connects intervention is important to consider because it provides evidence about the efficacy of comprehensive student support models to improve early and middle childhood academic outcomes, which are a major determinant of future educational attainment, success and health (Huurre, Aro, Rahkonen, & Komulainen, 2007).

The data source for the program evaluation portion of this dissertation research was a matched sample linking student demographic and state test data to District Z applicants' school lottery records. School lottery records were obtained directly from District Z and contain for each student their Student ID number, randomly generated lottery number, school preference list, and priority ranking for each school s/he ranked. Additionally, school lottery records included lottery application year (i.e., cohort), school choice received (i.e., lottery offer), and the priority ranking of the student for the school to which they received an offer. Student demographic and state test outcomes data were also obtained directly from District Z and were cleaned and re-structured into long-format to include a row for each year a student attended a school within District Z. The District Z files also contained information on student characteristics, such as race and gender, student academic performance measures, such as GPA and scores on the state assessment in English language arts and mathematics, and City Connects treatment variables indicating whether or not the student participated in the City Connects intervention and the number of years of the intervention received (i.e., dosage). Analytic files were subsequently created by merging student lottery records with the District Z student demographic and outcomes file based on students' unique Student ID number. Only those students found in both the lottery records and District Z student files were retained for analyses.

As District Z keeps records of every year for each student within the district, the District Z student file contains data spanning across multiple years of the City Connects implementation. However, for the purposes of this program evaluation, samples were limited to include only those students applying to District Z schools via lottery in the years 2006-2013. These years correspond to when District Z implemented the student-proposing deferred acceptance algorithm based on student-submitted school preference lists. Consequently, these data allowed the researcher to exploit the randomization within the District Z school assignment mechanism to conduct a natural experiment for estimating the impacts of City Connects on student outcomes. Subsequently, instrument-free methods were also applied to these same lottery data to illustrate alternative methods for dealing with endogeneity bias in estimating the City Connects treatment effect.

The City Connects intervention. Children growing up in poverty face a number of out-of-school factors that may impede their academic success and thriving, such as high rates of mobility, lack of quality healthcare, limited after-school or summer enrichment activities, and food insecurity (Berliner, 2009). Amidst growing evidence that out-of-school factors can affect academic outcomes, the City Connects intervention was developed to mitigate such barriers to learning through a systematic and coordinated process. As part of the City Connects evidence- and school-based intervention, a full-time

Master's level school counselor or social worker (the Coordinator) works with teachers to identify the strengths and needs of every student in their class across four developmental domains: academic, social/emotional/behavioral, health, and family. Guided by this information, the Coordinator matches each student to a tailored set of services through leveraging community resources, documents each student's service plan, and follows up to ensure the delivery of services (Walsh et al., 2014; Lee-St. John, 2012).

Developed in 2001 through a collaboration between Boston College, District Z, and area community agencies, the City Connects intervention has grown organically, adding schools and students by invitation and funding (Shields, Walsh, & Lee- St. John, 2016; Walsh, Raczek, Sibley, Lee- St. John, An, Akbayin, Dearing, & Foley, 2015). As a result, schools have not been randomly selected to participate in the City Connects intervention, and thus implementing a cluster randomized controlled trial design is impractical (Walsh et al., 2015). Evaluation of the City Connects treatment is further complicated by the fact that students do not randomly choose which school they attend within the District Z system, and instead the choice to attend a District Z school is determined by a myriad of non-random factors, such as socio-economic status and neighborhood. As a result, any study examining the effects of the City Connects intervention model will need to address endogeneity selection bias.

Lottery study design. Although City Connects is being implemented in five states and 84 schools as of 2019, the proposed program evaluation will be conducted on the implementation of the intervention in only District Z spanning the years 2006-2013. This timeline and site were selected because students within the District Z school district were assigned to schools via a centralized assignment system (Abdulkadiroglu et al., 2017)[1]. Within this system, students submit a preference list of schools, which asks students and their families to rank order up to ten schools in order of most preferred to least preferred. Each student is then assigned a priority ranking at each school to which they apply and, subsequently, students are rank ordered within schools based on these priorities, which can vary across schools. In the particular instance of the District Z school lottery system, students apply to programs within schools, known as "buckets." Furthermore, the District Z algorithm splits each bucket into two categories: a "walk" half, giving additional preference to walk zone applicants, and an "open" half, which gives no additional preference to applicants within a school's designated walk-zone. The priority rankings in hierarchical order of most preferred to least preferred are as follows:

Table 1.

Priority Rankings

Walk Slots	***Open Slots***
Guaranteed	Guaranteed
Sibling-Walk	
Sibling	Sibling
Walk	
No Priority	No Priority

Where *Guaranteed* means a student is granted automatic admission into the school s/he is applying to; *Sibling-Walk* indicates that a student has a sibling at the school to which s/he is seeking admission *and* that same student lives within that same school's

[1] In the years following 2012-13, District Z still relied on a centralized assignment mechanism for assigning students to schools; however, the mechanism was substantially changed, with walk-zones and school preference lists being eliminated from the process. Instead, a behind-the-scenes algorithm automatically populated a choice menu of schools for students to choose from based on zip code, thereby limiting the number of schools students could choose. This introduces additional complexities, therefore the researcher omitted any years after 2012-13 from the analysis.

designated walk zone (e.g., one mile radius); *Sibling* indicates that the student has a sibling already attending the school to which s/he is applying; *Walk* signifies that the student lives within a school's designated walk-zone; and *No Priority* means a student is given no preference consideration at the school to which s/he is applying. Once these priority rankings are established, the algorithm was implemented according to a walk-open precedence, such that students apply for a program's walk slot before applying to the same program's open half.

Given that each program/school has finite supply - i.e., a limited number of available seats - and students can be, and oftentimes are, tied on priority ranking within a given school, a decision rule is invoked to ration seats among students tied in ranking at each school. For the District Z school assignment mechanism, this decision rule is based on a single randomly generated lottery number that is completely independent of student priority and preferences. The rule is as follows: for any students who are applying to the same District Z school and tied on priority ranking within that school, the student(s) with the lowest lottery number(s) gains admission. In sum, the District Z algorithm considers the union of students' priority rankings and lottery number to form a strict rank ordering of students within each school. This feature of the assignment process generates valuable data for program evaluation researchers, as although assignment of priority rankings at each school are obviously non-random, reliance upon randomly generated lottery numbers for breaking ties on priority ranking creates conditional random assignment.

Given information contained in the school preference list, students' school-specific priority rankings, and the randomly generated lottery numbers, the District Z algorithm matches students to a District Z school program in an iterative fashion, terminating only when there are no longer any students applying to a school for which they have not yet been considered (students may be left unassigned). Specifically, the District Z deferred acceptance (DA) algorithm based on single tie-breaking operates as follows:

a) A single independently and identically distributed lottery number is drawn from a uniform distribution for each student.
b) Each student applies to his or her most preferred school, with each school rank ordering each of their applicants based upon priority ranking and lottery number combined.
c) Based upon this rank ordering of students, each school provisionally admits its highest ranked students up to the number of seats available at that school.
d) Each student rejected then goes on to apply to their next most applied school and competes with the highest ranked students provisionally admitted.

As outlined above, the District Z algorithm begins by matching students to their most preferred school program and for each program fills slots in order of student priority ranking and lottery number until capacity is reached. Subsequently, any student rejected from their top choice program in the previous step will go on to apply to their next most preferred school program, thereby competing with the pool of applicants with the lowest priority rankings and lottery numbers from the previous step. So for example, assume for simplicity that there are two students, denoted Student A and Student B, and two schools, labeled School One and School Two. Given that Student A has a sibling at School One, s/he then falls into the sibling priority category at that school. Meanwhile, assume Student B lives within a one mile radius of School One, and thus s/he then falls into the walk priority category at that school. Then, given that Student A and B both apply to School One and the sorting rules dictate that School One prefers students with sibling priority over students with walk priority, Student A is ranked and admitted at School One before Student B. If School One has capacity of one, and we change the hypothetical such that Student A and B are tied on priority ranking (e.g., both have walk status), then the student with the lower lottery number is granted admission. This random assignment feature serves as the foundation for IV approaches. Moreover, because City Connects is an intervention implemented within the District Z school district, the researcher can exploit the randomization within this District Z school assignment mechanism to obtain unbiased treatment effects of the City Connects intervention. As such, the implementation of City Connects in District Z provides an ideal opportunity to address the research question.

Methods and Data Analysis

This section addresses how the data were used to answer the research questions that guided this dissertation research. For each of the research questions, the specific analysis procedures will be discussed.

Analyses for Research Question 1

The first research question asks: *Under a range of endogeneity conditions, how does estimation performance under the two-stage least squares IV (2SLS-IV) approach, the Latent Instrumental Variable (LIV) approach, and the least squares Gaussian copula approach compare?* To address this question, simulation data (*data source 1*) were generated so that the following six studies could be conducted under various conditions:

Study 1. The researcher first generated data from a linear regression model where the assumptions of OLS were satisfied. Thus, endogeneity, denoted $\rho_{\varepsilon,x}$, was not present for this study.

Study 2. Data were generated from a linear regression model where enodogeneity was present and thus $\rho_{\varepsilon,x} \neq 0$; furthermore, endogeneity was varied such that $\rho_{\varepsilon,x}$ = 0.10, = 0.50, and 0.70.

Study 3. Data were generated from a Latent Instrumental Variable regression model where enodogeneity was present, $\rho_{\varepsilon,x} = 0.70$; furthermore, the Latent Instrumental Variable model varied in complexity, with three models specified: *a.)* a regression-through-the-origin model; *b.)* a model with an intercept and single slope; and *c.)* a model with intercept and multiple slopes. The distribution of the endogenous regressor was varied such that it took on a symmetric bimodal distribution and asymmetric bimodal distribution.

Study 4. Data were generated from both a Latent Instrumental Variable regression and linear regression model where endogeneity was present; for the LIV specification, $\rho_{\varepsilon,x} = 0.70$, and for the linear regression specification $\rho_{\varepsilon,x}$ =.10, =.30, and .70.

Study 5. Endogeneity was simulated from both a linear regression and Latent Instrumental Variable regression model. However, this time, the structural error distribution was specified to be non-normal for these data generating processes. Furthermore, the latent instrument and first-stage error were misspecified for the Latent Instrumental Variable regression model.

Study 6. Endogeneity was once again simulated from both linear regression and Latent Instrumental variable regression models. However, the sample size now varied for each specification, with N = 50, = 100, = 250, = 1000, = 2500, = 5000, representing small to considerably large samples. Additionally, the quality of the observed instrument for the instrumental variable analyses conducted for this study was varied such that the instrument took on the following four specifications: *a.)* high quality instrument; *b.)* weak but valid instrument; *c.)* strong, invalid instrument; *d.)* weak, invalid instrument.

For each condition across Studies 1-3, 5, and 6, the researcher generated 500 data sets, fitting OLS, instrumental variable, and instrument-free methods to each. For Study 4, the researcher generated 250 data sets due to the computational complexity of the model. Model estimates from the 250-500 data sets were then used to construct an empirical sampling distribution from which summary statistics could be calculated. For Studies 1-5, the researcher calculated the ratio of the distance of the mean estimated value from the true parameter value to the standard deviation of the estimated values, denoted t_{bias}, in order to establish unbiasedness of model estimates. For Study 6, the statistic of focus was mean squared error (MSE), which was calculated as follows:

$$MSE = E[(\hat{\beta} - \beta)^2] = bias^2 + \ variance \qquad (44)$$

The MSE of the different methods were then compared, with lower MSE values indicating better estimation performance. An outline of the various conditions for the six studies are shown in Table 2.

Table 2.

Overview of simulation studies

	Data Generation Process	**Goal of Study**	**Sample Size**	**Number of Simulations**	**Statistic(s) of Focus**
Study 1	Linear regression model; exogeneity	Investigate performance of methods under exogeneity	500	500	*t*-bias
Study 2	Linear regression model; endogeneity	Examine performance of methods under endogeneity arising from regression model	500	500	*t*-bias
Study 3	LIV model; endogeneity	Examine performance of methods under endogeneity arising from LIV model	500	500	*t*-bias
Study 4	Linear regression and LIV model; endogeneity	To further investigate performance of instrument-free methods under different endogeneity specifications	500	250	*t*-bias, correlation coefficient ρ
Study 5	Linear regression and LIV model; various error misspecifications	Examine robustness of methods to misspecification of error term	500	500	Bias; *t*-bias

Table 2 (continued).

Overview of simulation studies

Study 6	Linear regression and LIV model; endogeneity	Examine impact of sample size on method performance	50, 100, 250, 500, 1000, 2500, 5000	500	MSE

Further specific details regarding each study and the summary statistics used are discussed in Chapter 4.

Analyses for Research Question 2

The second research question asks the following: *Using data from a real-world school lottery study examining the effect of the City Connects model of integrated student support program, how do treatment effect estimates compare under the traditional 2SLS-IV approach with simulation-based propensity scores, the Latent Instrumental Variable (LIV) approach, and the Gaussian copula approach? And, how do the model parameters generated by instrument-free approaches compare to the observed instrument?* Data collected to evaluate the effects of the City Connects intervention on academic outcomes was used to address this research question. Specifically, the study used school lottery records, student demographic data, and student outcome data on the state test in English language arts and mathematics, each of which were obtained directly from District Z (*data source 2*). From these District Z supplied data, the researcher created two program evaluation data sets based on student participation in the District Z school lottery. The first evaluation data set (*data 2.a*) contained all District Z students from the lottery file for which demographic and student outcome data are also available, regardless of whether a student chose to opt out of the District Z school assignment process or participate[2]. Subsequently, a reduced data set was created by selecting only those students from the District Z lottery file who chose to submit a school preference list and thereby participate in the District Z school assignment process; this data set was then further refined to capture only those students who were assigned to a District Z school via randomization (*data 2.b*). A discussion of the specific details of this refinement process follows.

An observational analysis of the full program evaluation data set (*data 2.a*), was first conducted, producing naive *OLS* estimates of the City Connects effect. The linear regression model included a City Connects treatment variable as the endogenous regressor. Subsequently, the researcher subsetted the data based on random variation within the student proposing DA algorithm, selecting only those students for which the DA propensity score lies within the interval (0,1) (*data 2.b*). To achieve this, the researcher first identified and selected out only those students for which a preference list of schools was submitted. Therefore, if a student did not participate in the District Z school assignment process, s/he was deleted from the data file.

Following this, the researcher recreated the District Z school assignment process by coding up the DA algorithm along with the necessary user-specified inputs (e.g., preference matrices) in the R programming language. Specifically, the author used the "matchingR" package (Tilly & Janetos, 2018) in R for running the DA algorithm within a custom wrapper function. After the algorithm and user-specified inputs had been coded,

[2] If a student does not submit a school preference list, they opt out and are thus assigned by school administrators.
This process of administrative assignment takes places after the District Z school assignment process has concluded and is influenced by a host of non-random factors.

the allocation system was simulated one time. For the purposes of this study, the researcher operationalizes a reasonable degree of accuracy for the simulation as being 95% accuracy, i.e., 95% of students are assigned to a school via simulation that exactly matched the school they are assigned to via actual District Z lottery. When a reasonable degree of accuracy was not achieved based upon a single run of the algorithm, the algorithm and user-specified inputs were refined and the simulation process was run again.

Once a reasonable degree of accuracy had been achieved, the algorithm was then run n times with n different sets of student lottery numbers randomly drawn from a uniform distribution without replacement, creating a distribution of assignment. Subsequently, the results of these runs were tabulated, and a probability of school assignment for each student was calculated in frequentist fashion. Specifically, $p_a(i)$ = probability that student i gets assigned to school a = frequency with which the occurrence takes place across the n runs. These probabilities of school assignment are more formally known as DA propensity scores (Abdulkadiroglu et al., 2017), and students with DA propensity scores strictly between (0,1) were retained for analysis. These DA propensity scores were then used as covariates in the first- and second- stage equations of the IV analysis to control for assignment risk and create random assignment. The instrumental variable for this analysis was a dummy-coded variable indicating a random lottery offer to attend a City Connects school, and this was used in a 2SLS-IV regression to obtain exogenous variation in the City Connects treatment variable.

The simulated DA propensity score IV approach was chosen because other traditional lottery IV regression methods result in weaker, more inaccurate instruments and reduced sample sizes. Contrastingly, the DA propensity score-based stratification method reduces the dimensionality of preference and priority conditioning and identifies the maximal set of applicants subjected to randomized assignment (Abdulkadiroglu et al., 2017). Because the maximal set of applicants subject to randomization is identified, this approach arguably provides the most relevant and valid observed instrument to compare the instrument-free methods against. Results from this comparison may then provide compelling empirical validity evidence of the proposed approaches.

As previously noted, analyses based on extracting the random variation from the student assignment process reduced the sample size, as only a portion of students were actually assigned to a District Z school in a randomized way and furthermore not all students partake in the District Z school assignment process and instead opt out. As a result, this may have introduced selection bias despite the use of instrumental variables. To further explore this possibility, an extension of the simple LIV model, a nonparametric Bayesian LIV model estimated via MCMC estimation, and a Gaussian copula model via a CF approach were both applied to the full data (*data 2.a*). Because one of the benefits of an instrument-free approach is that no observed instrument is required, the data need not be subset based on a partially random mechanism, and thus the LIV and Gaussian copula approach can be used with the full data set used for OLS analysis, something that is not feasible when taking the traditional IV approach in lottery studies. The full data set is referred to as a quasi-lottery study, as it includes randomized participants along with non-randomized participants and non-participants. To then compare estimates of the City Connects causal effect across lottery study designs, both the Bayesian LIV model and Gaussian copula model were applied to the reduced data set (*data 2.b*), which reflects the lottery binding sample. Following this, City Connects causal effects were qualitatively compared across lottery study designs for evidence of lottery selection bias. Additionally, the researcher used this empirical application to further compare the instrument-free approaches with the traditional IV approach, applying the 2SLS-IV model to the same reduced data set as was used for the instrument-free approaches (*data 2.a*). Given that the data set used for 2SLS-IV, LIV, and the Gaussian copula is identical, direct comparisons of the estimated regression coefficients can provide useful insights. Thus, all resulting estimates and their associated precision were qualitatively evaluated and the causal effects of the City Connects treatment model were explored for interpretability across methods.

Lastly, a useful feature of the instrument-free model framework is that it allows for direct tests for exogeneity of the regressor using no observed instrument. For the LIV model, this test is calculated using the parameter estimates from the LIV model, $\hat{\beta}_{LIV}$ (Ebbes, 2004). More specifically, the researcher tested for exogeneity by assessing the 95% Credible Interval associated with the nonparametric Bayes LIV parameter estimate capturing the dependence between the endogenous regressor and the structural error term, denoted ρ.

The endogeneity test for the Gaussian copula method, the Hausman test, is simply a t-test on the regression coefficient associated with the copula Control Function term, X^* (Papies, Ebbes, & van Heerde, 2017). This was calculated as follows:

$$\frac{\hat{\beta}_{X^*}}{\sqrt{Var(\hat{\beta}_{X^*})}} \quad (45)$$

The researcher then examined the degree to which endogeneity presented as a problem across models, providing evidence for the appropriateness of IV methods in a lottery-study.

To address the second part of Research Question 2, *how the model parameters generated by these approaches compare to the observed instrument,* the researcher developed an optimal LIV instrument, denoted $\tilde{z}$, by fitting a nonparametric Bayesian LIV model and sampling from the conditional posterior distribution for the latent instrument. The optimal Bayes LIV instrument, $\tilde{z}$, was then calculated as the mean of this posterior distribution, rounded to the nearest integer (Ebbes, 2004).Once this was achieved, the researcher compared the optimal LIV instrument with the observed instrument for the traditional 2SLS-IV lottery analysis, examining classifications across methods via a contingency table. Furthermore, The City Connects treatment variable was correlated with the optimal LIV instrument to assess the relevance of the optimal LIV instrument, and a 2SLS estimation of the City Connects treatment effect using the optimal LIV instrument was compared to the treatment effect estimate obtained from using the observed lottery instrument. Lastly, the 2SLS-LIV regression estimate was also compared to the Gaussian copula City Connects effect estimate.

CHAPTER 4: ANALYSES AND RESULTS

Research Question One

The first research question aimed to compare the performance of the LIV and Gaussian copula instrument-free methods with ordinary least squares and instrumental variable regression methods across a range of research conditions. In total, six discrete studies were conducted to address this research question. Study 1 involved the formulation of a baseline model with no endogeneity. Studies 2 and 3 investigated performance of OLS, IV, and instrument-free methods under endogeneity arising from both a linear regression and LIV model. Study 4 then further explored the impact of different endogeneity specifications on instrument-free method performance. Study 5 explored misspecification of the error distribution and Study 6 focused on the impact of sample size and instrument quality on the performance of instrumental variable and instrument-free methods. Table 2 presents a summary of the studies conducted by describing the data generation processes, the goal of the study, the sample size used, the number of simulations, and the statistics that were used to evaluate the methods.

Analyses were performed using R and WinBUGS software; specifically, the author used the "AER" package (Kleiber & Zeileis, 2008) and code adapted from the "REndo" package (Gui, Meirer, Algesheimer, & Schilter, 2019) in R. All syntax used is available upon request. In the following sections, the analyses and results for each of the six studies explored for this dissertation research are presented and discussed.

Study 1: Exogeneity

The researcher first investigated the performance of the LIV and Gaussian copula regression methods under the research condition of exogeneity. This research condition served as a baseline, comparing instrument-free methods with ordinary least squares under ideal experimental conditions. The data generating process (DGP) was specified to be the classic linear regression model as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 0.8 + 0.7 x_i + \varepsilon_i, \quad (46)$$

$$\varepsilon \sim N(0, \sigma^2) = N(0, 1),$$

For this specification no endogeneity was present such that $\rho_{x,e} = 0$. Additionally, x_i was fixed and discrete and the sample size, N, was set to 500. Unless otherwise noted, a sample size of $N = 500$ remained constant throughout this dissertation research until Study 6. Five-hundred hundred data sets of N = 500 were generated under the above specified conditions and ordinary least squares, latent instrumental variable, and Gaussian copula regression models were fitted. The model estimates obtained across the 500 data sets were subsequently used to construct the empirical sampling distributions of the parameter estimates β_0 and β_1 (Park & Gupta, 2012). The researcher investigated the means and standard deviations of the empirical sampling distribution and calculated t_{bias} to make inferences about the difference between the mean estimates and the true parameter value. The t_{bias} statistic is the ratio of the distance of the mean estimated value from the true parameter value to the standard deviation of the estimated values and allowed the researcher to establish unbiasedness[3] . As is standard in the literature, the researcher used a cut point of 1.96 to establish unbiasedness using the t_{bias} statistic, as this represents roughly two standard errors and is the value corresponding to the critical *t*-

[3] In actuality, the author is referring to asymptotic unbiasedness, as IV and instrument-free methods are unbiased in the limit; furthermore, the LIV estimator is approximately consistent, as consistency of the estimator has only been shown via simulation and no formal mathematical proof for consistency has been given to date.

statistic for a 95% confidence interval (Park & Gupta, 2012). Means, standard errors, and t_{bias} are provided in Table 3.

Given that the true values for the intercept and regression coefficient were 0.80 and 0.70, respectively, we see from the mean and standard error estimates in Table 3 that the OLS model produces the best linear unbiased estimate of β_0 and β_1. This is to be expected, as the assumptions of OLS, in particular for exogeneity, have been satisfied under this particular simulation study design. Interestingly, however, both the LIV and Gaussian copula regression models also yield unbiased estimates of β_0 and β_1, even when no endogeneity is present. By examining the standard errors in Table 3, we note that the LIV estimate for the main parameter of interest, β_1, is the least efficient, with standard errors that are more than twice that of the OLS estimate.

Study 2: Endogeneity from Linear Regression Model

Study 2 investigated the performance of instrument-free methods compared with traditional instrumental variable regression under a range of endogeneity conditions from minor to severe, where $\rho_{x,e}$ = 0.10, = 0.30, and 0.50. The DGP and parameter values used were as follows:

$$\begin{pmatrix} x_i^* \\ \varepsilon_i^* \\ z_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{x,e} & 0.5 \\ \rho_{x,e} & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right),$$

$$x_i = F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\ \Phi(x_i^*)\ ) = F_{x,i}^{-1}(\ \Phi(x_i^*)\ | a, \beta\ ),$$

$$\varepsilon_i = F_{e,i}^{-1}(U_{\varepsilon,i}) = F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) = \Phi^{-1}(\ \Phi(\varepsilon_i^*)\ ),$$

$$z_i = U_{z,i} = \Phi(z_i^*)\ ,$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5x_i + \varepsilon_i, \ (47)$$

where $F_{x,i}^{-1}(\ |\alpha,\beta)$ is the inverse cumulative gamma distribution with shape and scale parameters α = 2 and β = 2, and Φ is the standard normal cumulative distribution. Therefore, x_i has a $Gamma(2,2)$ distribution, ε_i follows the standard normal distribution, and the true instrument, z_i, follows a continuous uniform distribution on the (0,1) interval. The true instrument, z_i, is exogenous and strongly correlated with the endogenous regressor (i.e., $\rho_{z,x}$ = 0.5). Note that the LIV model is misspecified for these conditions, as the endogenous regressor is not structurally composed of a discrete, exogenous instrument and an additive endogenous disturbance term. Instead, a continuous true instrument is correlated with the endogenous regressor.

For each endogeneity condition, $\rho_{x,e}$ = 0.10, = 0.30, and 0.50, OLS, IV, LIV, and Gaussian copula regression models were fit to the 500 data sets generated. Model estimates from across the 500 data sets were used to construct the empirical sampling distributions of β_0 and β_1 (Park & Gupta, 2012). Table 4 shows the means, standard errors, and t_{bias} under each condition.

Even for the mild endogeneity condition where $\rho_{x,e}$ = 0.10, we see from the t_{bias} statistics that OLS produced biased results for the regression coefficient associated with the endogenous regressor. Furthermore, as endogeneity increases, we see that OLS estimates become increasingly biased, producing mean estimates that deviate considerably from the true parameter values. For example, under the endogeneity condition of $\rho_{x,\epsilon} = 0.5$ where the true parameter value is 2.5, the OLS model produces a mean estimate of 3.17. Unsurprisingly, IV estimates are unbiased regardless of the level of endogeneity specified. This is to be expected, as the IV model makes use of the true instrument, which is strongly correlated with the endogenous regressor and uncorrelated with the structural error term; as a result, the instrument used is ideal. For the instrument-free methods, we see that the Gaussian copula estimates are very close in value to the true parameter values with bias that is not significantly different from zero; additionally, the LIV estimates are also unbiased, albeit less accurate than the Gaussian copula estimate[4]. The benefit is that instrument-free methods provide unbiased results without relying on the true instrument. However, we note that the LIV estimates may only be unbiased due to the inefficiency of the method under the linear regression specification, as mean estimates for the approach notably deviate from the true parameter value, especially for larger values of endogeneity.

Although the estimate from the OLS model is biased, we see that it remains the most efficient, yielding standard errors that are smaller than those from all other specified methods. This is problematic, however, as the distribution of the estimates is tightly distributed around an erroneous and biased value. Consistent with the extant literature on instrumental variables (Ebbes, 2004; 2009; Boef, Dekkers, Vandenbroucke, & le Cessie, 2014), the IV estimate is less efficient than OLS, yielding a standard error nearly double in size. Notably, the instrument-free methods yield the largest standard errors, indicating less precision. The reduced efficiency compared to 2SLS IV regression is due to the fact that the IV approach uses the true, ideal instrument. Moreover, there is an additional efficiency cost for the LIV approach due to misspecification (Ebbes et al., 2009).

[4] For maximization of the likelihood function, the author used a derivative-free optimization technique. If one opts for a quasi-newton approach, it is advised that analytic expressions of the gradient and Hessian be used, as more stable and accurate results will likely be obtained by relying on such expressions. This is because when the gradient and Hessian information are not given, numerical approximations of them must be calculated, and such approximations introduce error.

Table 3.

Study 1: Exogeneity

		OLS			LIV			Gaussian copula		
Θ	**True Value**	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}
β_0	0.80	0.80	0.11	0.04	0.91	0.25	0.45	0.81	0.17	0.06
β_1	0.70	0.70	0.03	-0.01	0.67	0.08	-0.45	0.70	0.07	-0.04

Table 4.

Linear Regression Endogeneity

			OLS			IV			LIV			Gaussian copula		
$\rho_{x,\epsilon}$	**Θ**	**True Value**	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}
0.1	β_0	1.00	0.86	0.08	-1.67	1.00	0.15	-0.03	1.09	0.27	0.34	1.00	0.22	-0.02
	β_1	2.50	2.64	0.07	2.10	2.51	0.14	0.04	2.42	0.27	-0.30	2.50	0.22	0.02
0.3	β_0	1.00	0.59	0.08	-5.31	1.00	0.15	0.01	0.94	0.25	-0.24	0.99	0.21	-0.04
	β_1	2.50	2.91	0.06	6.47	2.50	0.14	-0.03	2.57	0.24	0.28	2.50	0.21	0.01
0.5	β_0	1.00	0.33	0.07	-9.65	1.00	0.15	0.01	0.86	0.32	-0.44	0.99	0.20	-0.03
	β_1	2.50	3.17	0.06	10.76	2.50	0.14	-0.01	2.65	0.31	0.49	2.50	0.19	0.01

Study 3: Endogeneity under LIV Model

For Study 3, endogeneity based on the LIV model was generated to further compare OLS, IV, LIV, and Gaussian copula methods. For this study, the endogenous regressor is structurally different than it was in Study 2, with it now being decomposed into two pieces: the true, discrete instrument with two categories, and an additive, endogenous Gaussian error term (Ebbes et al., 2009). The researcher simulated data from three DGP's with varying parameters to compare estimation performance. These studies are described as Studies 3.1 to 3.3 in the sections that follow.

Study 3.1: Regression through the origin (RTO) LIV model. For this sub-study, the DGP's and parameters were specified using a design similar to that appearing in the 2012 work by Park and Gupta, where the authors compared the performance of the Gaussian copula model with the latent instrumental variable model under the RTO LIV research design. Thus, the researcher used this first DGP as a baseline for providing comparative evidence.

$$\begin{pmatrix} \nu_i \\ \varepsilon_i \\ z_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right),$$

$$z_i = F_{z,i}^{-1}(U_{z,i}) = F_{z,i}^{-1}(\ \Phi(z_i^*)\ ) = F_{z,i}^{-1}(\ \Phi(z_i^*)\ , n, p_z\ ),$$

$$x_i = z_i\pi + \nu_i\ = \begin{cases} F_{x,i}^{-1}\left(U_{x,i}\right) = \ F_{x,i}^{-1}(\ \Phi(\nu_i)\ ) = \ \Phi_{(6,1)}^{-1}(\ \Phi(\nu_i)\ ),\ if\ z_i = 0 \\ F_{x,i}^{-1}\left(U_{x,i}\right) = \ F_{x,i}^{-1}(\ \Phi(\nu_i)\ ) = \ \Phi_{(2,1)}^{-1}(\ \Phi(\nu_i)\ ),\ if\ z_i = 1 \end{cases},$$

$$y_i = \beta_1 x_i + \ \varepsilon_i = 2.5x_i + \ \varepsilon_i,\ (48)$$

Study 3.2: Full LIV model with intercept. The specifications of this model are as before, however, an RTO model is a very simple parameterization that is rarely used in practice, as it can lead to biases when inappropriate. Therefore, the researcher extended the DGP to investigate the performance of these methods under a fuller parameterization that included an intercept and slope as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5x_i + \varepsilon_i, \quad (49)$$

Study 3.3: Full LIV model with additional exogenous regressor. The researcher then extended the DGP further to include an additional control variable, investigating the performance of the methods under a fuller parameterization that included an intercept and multiple slopes.

$$\begin{pmatrix} v_i \\ \varepsilon_i \\ z_i^* \\ c_i \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0 & 0.1 \\ 0.7 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.1 & 0 & 0 & 1 \end{bmatrix} \right),$$

$$z_i = F_{z,i}^{-1}(U_{z,i}) = F_{z,i}^{-1}(\,\Phi(z_i^*)\,) = F_{z,i}^{-1}(\,\Phi(z_i^*)\,, n, p_z\,),$$

$$x_i = z_i \pi + v_i = \begin{Bmatrix} F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\,\Phi(v_i)\,) = \Phi_{(6,1)}^{-1}(\,\Phi(v_i)\,), \ if\ z_i = 0 \\ F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\,\Phi(v_i)\,) = \Phi_{(2,1)}^{-1}(\,\Phi(v_i)\,), \ if\ z_i = 1 \end{Bmatrix},$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 c_i + \varepsilon_i = 1 + 2.5x_i + 0.8c_i + \varepsilon_i, \quad (50)$$

where $\Phi_{(6,1)}^{-1}$ is the inverse normal distribution function with $\mu = 6$ and $\sigma^2 = 1$; $\Phi_{(2,1)}^{-1}$ is the inverse normal distribution function with $\mu = 2$ and $\sigma^2 = 1$; and $F_{z,i}^{-1}$ is the inverse binomial cumulative distribution function with number of trials $n = 1$ and probability of success for each trial p_z = 0.5, 0.8. Therefore, we note that $\pi = (6, 2)'$ with probabilities p_z and $(1 - p_z)$. Furthermore, by varying p_z, the researcher changes the distribution of x_i; when p_z =0.5, x_i is bimodal and symmetric with equal maxima and when p_z =0.8, x_i is bimodal with unequal maxima. Thus, in general, $x_i \sim p_z * N(6,1) + (1 - p_z) * N(2,1)$. The endogenous regressor, x_i , will be correlated with the structural error by roughly $\frac{1}{2} * \rho_{\varepsilon,v}$, given the specification. Given that x_i comprises a discrete, exogenous instrument and an endogenous Gaussian disturbance term, ν_i, the LIV model is correctly specified for these conditions. However, we note that the Gaussian copula assumes a different dependence structure than the one specified, namely it assumes that the dependence between the error term and the endogenous regressor follows a copula dependence structure, and thus it is misspecified for these conditions (Park & Gupta, 2012).

For Study 3.3 (i.e., the full LIV model with an additional exogenous variable), a nonparametric Bayesian LIV model was estimated instead of the simple LIV model used with Studies 3.1 and 3.2, as it allows for the inclusion of additional exogenous regressors in a simple manner and avoids identification issues that may arise with the traditional MLE approach (Ebbes, 2004). In contrast to the simple LIV model, the Bayesian LIV model does not impose restrictions on the distribution of the latent instrument but instead specifies a Dirichlet process as a prior distribution on the space of all possible distribution functions for the latent instrument. Simply put, a Dirichlet process is a collection of random variables, where each random variable is itself a probability distribution function; therefore, one can think of it as a distribution over distributions, and with each draw from the distribution also yielding a probability distribution (Whye Teh, 2010). Fundamental to Bayesian inference is that we assign prior distributions to the unknown quantities in a model (Gorur & Rasmussen, 2010). However, in order to do so, we must identify the parameters of the prior distribution (i.e., know its parametric form). Using a Dirichlet process as a nonparametric prior allows us to express any uncertainty about this parametric form (Gorur & Rasmussen, 2010). Given the Dirichlet process prior specification for the Bayesian LIV model, the unknown distribution for the latent instrument is then estimated from the data (Ebbes, 2004; 2005). The Bayesian LIV model takes the following form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{(51A)}$$

$$x_i = \theta_i + v_i, \quad \text{(51B)}$$

$$where\ \theta_i\ are\ independently\ drawn\ from\ G\ and$$

$$G \sim DP(\alpha, G_0)$$

Unlike the simple LIV model, where we assumed a multinomial distribution with m outcomes for the latent instrument, we now make no parametric assumptions regarding the form of the latent instrument distribution (Ebbes, 2004; Ebbes, Bockenholt, Wedel, Nam, 2014). The distribution G is then given a Dirichlet process prior, $DP(\alpha, G_0)$, with non-negative concentration parameter α and baseline prior distribution G_0 (Ebbes, 2004; Ebbes et al., 2014). Given the added computational complexity of estimating the Bayesian LIV model, the number of simulations for Study 3.3 was set to 250. For Studies 3.1 and 3.2, the number of simulations was set to 500. Tables 5, 6, and 7 show the results for OLS, IV, LIV, nonparametric Bayesian LIV, and Gaussian copula regression models.

For data generated from an intercept-only LIV model, we see that OLS produces upwardly biased estimates ($\hat{\beta}^{OLS} = 2.53, = 2.56$), as mean estimates from this approach are statistically significantly larger than the true parameter value. Given the endogeneity problem, this is to be expected; however, the instrumental variable regression approach, which relies on the true, ideal instrument, corrects for this bias, producing accurate, unbiased estimates for the slope, β_1 that recapture the true parameter values. Both instrument-free methods, which do not rely on an observed instrument, also produce accurate, unbiased results. Furthermore, the Gaussian copula approach produces unbiased results despite misspecification, which matches simulation results reported in Park and Gupta (2012). This result is only for a simple parameterization, however, and we note that when endogeneity was based on a fully specified model with both intercept and slopes, the Gaussian copula estimates for both β_0 and β_1 became significantly biased, deviating considerably from the true parameter values[5]. Moreover, this bias occurred for both a symmetric and asymmetric endogenous regressor, with the copula estimate significantly underestimating the true causal effect. Conversely, both the LIV and nonparametric Bayesian LIV approaches produce unbiased, highly accurate results regardless of the parameterization specified. Such results suggest that the dependence structure matters, and the Gaussian copula approach has difficulty adapting to endogeneity processes that differ from that which is assumed for the model.

[5] The biased intercept estimate given by the Gaussian copula approach can be addressed by mean-centering the regressor; however, this strategy was investigated and it did not change the result of the slope for the endogenous regressor being significantly biased. Moreover, centering approaches did not change the results across all specified research conditions.

Table 5.

RTO LIV Model

			OLS			IV			LIV			Gaussian copula		
p_z	Θ	True Value	Mean	S.E.	t_{bias}	Mean	S.E.	t_{bias}	Mean	S.E.	t_{bias}	Mean	S.E.	t_{bias}
0.5	β_1	2.50	2.53	0.01	3.50	2.50	0.001	-0.04	2.50	0.03	-0.07	2.49	0.01	-1.19
0.8	β_1	2.50	2.56	0.01	4.90	2.50	0.01	0.03	2.50	0.03	-0.05	2.48	0.02	-1.39

Table 6.

LIV Model w/ Intercept

			OLS			IV			LIV			Gaussian copula		
p_z	Θ	True Value	Mean	S.E.	t_{bias}	Mean	S.E.	t_{bias}	Mean	S.E.	t_{bias}	Mean	S.E.	t_{bias}
0.5	β_0	1.00	0.44	0.08	-7.12	1.01	0.09	0.06	1.02	0.11	0.14	3.20	0.35	6.21
	β_1	2.50	2.64	0.02	8.34	2.50	0.02	-0.06	2.50	0.03	-0.08	1.95	0.09	-6.33
0.8	β_0	1.00	0.45	0.07	-7.65	1.00	0.10	0.04	1.02	0.10	0.15	2.52	0.27	5.63
	β_1	2.50	2.70	0.02	8.64	2.50	0.03	-0.01	2.50	0.03	-0.08	1.95	0.09	-5.83

Table 7.

LIV Model w/ Additional Exogenous Regressor

			OLS			IV			Bayesian LIV			Gaussian copula		
p_z	Θ	**True Value**	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}
0.5	β_0	1.00	0.44	0.08	-6.76	1.00	0.09	-0.01	0.98	0.17	-0.10	3.21	0.35	6.34
	β_1	2.50	2.64	0.02	7.88	2.50	0.02	0.04	2.50	0.04	0.11	1.94	0.09	-6.41
	β_2	0.80	0.78	0.04	-0.40	0.80	0.04	-0.09	0.80	0.04	-0.09	0.76	0.04	-1.13
0.8	β_0	1.00	0.44	0.07	-7.91	0.99	0.09	-0.06	0.98	0.20	-0.12	2.52	0.28	5.38
	β_1	2.50	2.70	0.02	8.67	2.50	0.03	0.03	2.51	0.07	0.11	1.95	0.10	-5.83
	β_2	0.80	0.78	0.04	-0.49	0.80	0.04	0.03	0.80	0.04	0.01	0.75	0.04	-1.21

Study 4: Investigating Endogeneity with an Optimal Bayes LIV Instrument

When taken collectively, evidence from Studies 2 and 3 strongly suggests that the dependence structure matters when choosing between the two instrument-free approaches. As shown in Study 3, when endogeneity was specified according to a full LIV model (i.e., there was additive separability in the endogeneous regressor and the model included both intercept and slopes) the least squares Gaussian copula method yielded significantly biased estimates. Conversely, when endogeneity was specified according to the linear regression model (i.e., endogeneity was purely correlational) the LIV method provided far less accurate, albeit statistically unbiased, parameter estimates than the Gaussian copula approach. This finding suggests a potential problem with the LIV approach under the linear regression endogeneity specification despite reported t_{bias} statistics being less than the 1.96 threshold. In other words, the researcher wonders whether the deviation of the LIV estimates from the true parameter value under a linear regression specification is an actual problem not being captured by the t_{bias} statistics. This is important to consider because if the LIV model is significantly biased under a linear regression specification, then the dependence structure matters to both the Gaussian copula and LIV model and an additional assumption is imposed upon the instrument-free approaches.

To further investigate the impact of the dependence structure, the researcher simulated data from two DGP's, each with a different endogeneity specification. The DGP's and true parameters are as follows:

DGP 4.1: Full LIV model endogeneity

$$\begin{pmatrix} \nu_i \\ \varepsilon_i \\ z_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right),$$

$$z_i = F_{z,i}^{-1}(U_{z,i}) = F_{z,i}^{-1}(\,\Phi(z_i^*)\,) = F_{z,i}^{-1}(\,\Phi(z_i^*)\,, n, p_z\,),$$

$$x_i = z_i\pi + \nu_i = \begin{cases} F_{x,i}^{-1}\left(U_{x,i}\right) = F_{x,i}^{-1}(\,\Phi(\nu_i)\,) = \Phi_{(6,1)}^{-1}(\,\Phi(\nu_i)\,),\ if\ z_i = 0 \\ F_{x,i}^{-1}\left(U_{x,i}\right) = F_{x,i}^{-1}(\,\Phi(\nu_i)\,) = \Phi_{(2,1)}^{-1}(\,\Phi(\nu_i)\,),\ if\ z_i = 1 \end{cases},$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5x_i + \varepsilon_i \quad (52)$$

DGP 4.2: Linear regression endogeneity

$$\begin{pmatrix} x_i^* \\ \varepsilon_i^* \\ z_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{x,e} & 0.5 \\ \rho_{x,e} & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} \right),$$

$$x_i = F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\,\Phi(x_i^*)\,) = F_{x,i}^{-1}(\,\Phi(x_i^*)\,|a, \beta),$$

$$\varepsilon_i = F_{e,i}^{-1}(U_{\varepsilon,i}) = F_{e,i}^{-1}(\,\Phi(\varepsilon_i^*)\,) = \Phi^{-1}(\,\Phi(\varepsilon_i^*)\,),$$

$$z_i = F_{z,i}^{-1}(U_{z,i}) = F_{z,i}^{-1}(\,\Phi(z_i^*)\,) = F_{z,i}^{-1}(\,\Phi(z_i^*)\,, n, p_z\,),$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5x_i + \varepsilon_i, \ (53)$$

where the specifications are the exact same as those in Studies 2 and 3, except now the true instrument, z_i, takes a binomial distribution across the two DGP's. The researcher then sampled 250 times from both *DGP 4.1* and *4.2*, fitting a nonparametric Bayesian LIV model from Equation 51 to each sampled data set from each DGP. In fitting the nonparametric Bayesian LIV model to the data, the researcher produced a posterior distribution for the optimal Bayes LIV instrument by sampling from the following full conditional distribution:

$$p(\tilde{z}_i \mid \tilde{z}_{-i}\, \beta, \Sigma, \alpha, G_0, b)\,, \qquad (54)$$

where b is the $n\,x\,1$ vector containing elements b_i, and where $b_i = (y_i, x_i)$[6]. The optimal Bayes LIV instrument, $\tilde{z}$, was then calculated as the mean of this posterior distribution, rounded to the nearest integer. The rounded posterior mean produces an observed instrument from the estimated Bayes LIV model, allowing for comparison of this instrument produced by the LIV model to the true, observed instrument, z. With the optimal Bayes LIV instrument, the researcher performed the following analyses across the two DGP's:

1) 2SLS-IV regression using the optimal Bayes LIV instrument, comparing the coefficients from this method to those from the Bayes LIV and 2SLS-IV regression using the true instrument;
2) correlation analysis of the optimal Bayes LIV instrument with the true instrument;
3) correlation analysis of the optimal Bayes LIV instrument with the endogenous regressor (*relevance*); and
4) correlation analysis of the optimal Bayes LIV instrument with the true error (*exogeneity*).

The researcher then examined results from the four analyses to investigate how well the LIV model recovered the true parameter values and reproduced the true instrument for both DGP's.

LIV endogeneity. Tables 8 and 9 provide results for data generated from *DGP 4.1,* which represents endogeneity arising from a LIV model. Specifically, Table 8

[6] See Ebbes 2004 for the conditional posterior distributions derived in full detail

provides model estimates from a 2SLS-IV regression with the optimal Bayes LIV instrument, which are referred to as 2SLS-LIV, a Bayes LIV regression, and a 2SLS-IV regression with the true instrument. Table 9 provides results from correlational analyses performed with the optimal Bayes LIV instrument.

From the t_{bias} statistics in Table 8, we see that OLS estimates statistically significantly differ from the true parameter values. Given the endogeneity problem, this is to be expected. Moreover, as previously demonstrated, both the IV regression and the Bayes LIV regression approaches provide unbiased results. Moreover, we also see that the 2SLS-LIV approach, which uses an optimal Bayes LIV instrument estimated from the data as the observed instrument, produces highly accurate, unbiased estimates of the true parameter value. Interestingly, the 2SLS-LIV approach is more efficient than the Bayesian LIV approach, yielding standard errors that are less than half the size of the standard errors from the Bayesian LIV model. In fact, the 2SLS-LIV standard errors are equal to those from the IV approach using the true observed instrument, suggesting that the optimal LIV instrument estimated from the data is equivalent to the true instrument. Thus, there is an efficiency gain to estimating the latent instrument and using it in a two-stage least squares regression.

In looking at Table 9, we see that the Bayes LIV model produces a highly accurate and valid instrument. The correlation between the optimal Bayes LIV instrument and the true instrument is .99, and the classification accuracy for the estimated instrument (i.e., how well the optimal Bayes LIV instrument captures true instrument group membership) is 99.7%. We also see from Table 10 that the optimal Bayes LIV instrument is highly relevant (i.e., $\bar{\rho}_{\tilde{z},x}$ = -.89, = -.85) and uncorrelated with the true error ($\bar{\rho}_{\tilde{z},\varepsilon} = <$

.01). In sum, results suggest that the LIV approach reproduces the true instrument very well, given that the dependence structure is what the method assumes.

Linear regression endogeneity. Tables 10 and 11 provide results for data generated from *DGP 4.b,* which represents endogeneity under a linear regression specification. In examining Table 10, we once again see that the LIV approach yields inaccurate yet statistically unbiased results across all specified endogeneity conditions ($\rho_{x,\epsilon} = 0.1$, = 0.3, and 0.5). This result matches the results seen in Study 2; however, in looking at the 2SLS-LIV results, we now see empirical evidence for this unbiasedness being solely due to the inefficiency of the LIV method under misspecification, as the 2SLS-LIV approach yields inaccurate and statistically biased results. Furthermore, Table 11 reveals that the optimal Bayes LIV instrument estimated from the data is no longer valid. We see that the optimal Bayes LIV instrument is weakly correlated with the true instrument ($\bar{\rho}_{\tilde{z},z} =$ -.22, = -.23) and the classification accuracy of 43% is poor. Furthermore, while the optimal Bayes LIV instrument is still highly relevant ($\bar{\rho}_{\tilde{z},x}$ = -.72, = -.71), it is now correlated with the true error ($\bar{\rho}_{\tilde{z},\varepsilon}$= - .05, = - .14, = - .23). Interestingly, the correlation between the optimal LIV instrument and the true error is about half of the endogeneity specified for the DGP. Thus, we see that the LIV approach under the linear regression specification essentially produces a less endogenous regressor, thereby producing biased estimates, albeit estimates closer to the true parameter value than those produced by OLS. In other words, there is residual endogeneous variation left over when the LIV model is fit to data generated from a linear regression specification and the dependence structure is different from that assumed by the model.

Table 8.

Endogeneity under DGP 4.1

			OLS			IV			Bayes LIV			2SLS-LIV		
p_z	Θ	True Value	Mean	S.E.	t_{bias}	Mean	S.E.	t_{bias}	Mean	S.E.	t_{bias}	Mean	S.E.	t_{bias}
0.5	β_0	1.00	0.44	0.09	-6.46	1.01	0.10	0.05	0.99	0.19	-0.03	1.00	0.10	0.05
	β_1	2.50	2.64	0.02	8.07	2.50	0.02	-0.03	2.50	0.05	0.05	2.50	0.02	-0.02
0.8	β_0	1.00	0.45	0.07	-7.73	1.00	0.09	0.04	0.95	0.31	-0.17	1.00	0.09	0.03
	β_1	2.50	2.70	0.02	8.80	2.50	0.03	-0.04	2.52	0.11	0.18	2.50	0.03	-0.03

Table 9.

Correlation analyses for DGP 4.1

Optimal Bayes LIV Instrument Correlations				
p_z	$\bar{\rho}_{\tilde{z},z}$	% correctly classified	$\bar{\rho}_{\tilde{z},x}$	$\bar{\rho}_{\tilde{z},\epsilon}$
0.5	0.99	99.7%	-0.89	< .001
0.8	0.99	99.8%	-0.85	< .001

Table 10.

Endogeneity under linear regression model DGP 4.2

			OLS			IV			Bayes LIV			2SLS-LIV		
$\rho_{x,\epsilon}$	**Θ**	**True Value**	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}	**Mean**	**S.E.**	t_{bias}
0.1	β_0	1.00	0.87	0.08	-1.65	1.02	0.17	0.12	0.90	0.35	-0.29	0.91	0.14	-0.65
	β_1	2.50	2.63	0.07	1.94	2.48	0.17	-0.13	2.60	0.34	0.29	2.59	0.13	0.67
0.3	β_0	1.00	0.60	0.08	-5.24	1.01	0.18	0.06	0.70	0.31	-0.96	0.73	0.13	-2.10
	β_1	2.50	2.90	0.06	6.31	2.49	0.17	-0.06	2.80	0.31	0.96	2.77	0.12	2.26
0.5	β_0	1.00	0.33	0.07	-9.38	1.01	0.17	0.05	0.58	0.32	-1.29	0.55	0.12	-3.80
	β_1	2.50	3.17	0.06	10.89	2.50	0.17	-0.02	2.92	0.33	1.29	2.96	0.11	4.14

Table 11.

Correlation analyses for DGP 4.2

Optimal Bayes LIV Instrument Correlations				
$\boldsymbol{\rho_{x,\epsilon}}$	$\boldsymbol{\bar{\rho}_{\tilde{z},z}}$	**% correctly classified**	$\boldsymbol{\bar{\rho}_{\tilde{z},x}}$	$\boldsymbol{\bar{\rho}_{\tilde{z},\epsilon}}$
0.1	-0.22	43%	-0.72	-0.05
0.3	-0.22	43%	-0.72	-0.14
0.5	-0.23	43%	-0.71	-0.23

Study 5: Error Misspecification

For Study 5, the researcher examined the robustness of instrument-free methods to misspecification of the error term. To do so, the researcher varied the error distribution across a range of different DGP's and examined model estimation across the different specifications. These studies are described as Studies 5.1 to 5.4 in the sections that follow. Each of the following subsections will outline the DGP and error distributions specified, true parameter values, and results.

Study 5.1: Exogeneity. Data was first simulated from a linear regression model with a non-normal error term and exogenous regressor. The DGP and parameter values generated were as follows:

$$\begin{pmatrix} x_i^* \\ \varepsilon_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right),$$

$$x_i = F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\ \Phi(x_i^*)\ ) = F_{x,i}^{-1}(\ \Phi(x_i^*)\ |\alpha,\beta\ ),$$

$$\varepsilon_i = \begin{cases} U_{\varepsilon,i}\ =\ \Phi(\varepsilon_i^*) - E(\ \Phi(\varepsilon_i^*)\ ), & Condition\ 1 \\ F_{e,i}^{-1}(U_{\varepsilon,i}) =\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) =\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ |\ \nu_1,\nu_2\ ) - E\big(\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ |\ \nu_1,\nu_2\ )\ \big), & Condition\ 2 \\ F_{e,i}^{-1}(U_{\varepsilon,i}) =\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) =\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ |\ k\ ) - E\big(\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ |\ k\ )\ \big), & Condition\ 3 \end{cases}$$

$$y_i = \beta_0 + \ \beta_1 x_i + \ \varepsilon_i = 1 + 2.5x_i + \ \varepsilon_i,\ (55)$$

where Φ is the standard normal cumulative distribution, $F_{e,i}^{-1}(\ \mid d_1, d_2\ )$ is the inverse cumulative *F*-distribution with parameters $d_1 = 8$ and $d_2 = 5$, $F_{e,i}^{-1}(\ \mid k\ )$ is the inverse cumulative chi-square distribution with parameter $k = 4$, and $E(\ )$ is the expectation operator. Thus, the error term takes on the uniform, *F*- and Chi-square distributions each with mean zero. Given the error specifications, all models are misspecified. Table 12 provides the results for each specification of the error term.

As shown by Table 12, OLS produces estimates of the intercept and causal effect with both the smallest bias and standard error (e.g., $Bias^{OLS}$ = >-.01; $S.E.^{OLS}$ = .02). Such a result is ensured by the Gauss-Markov theorem, where the assumption of normality is not required for proving unbiasedness and efficiency of the OLS estimator[7]. Interestingly, the Gaussian copula approach also gives very accurate results. Given the least squares specification of the model and the absence of endogeneity, this result is as expected, as the distributional assumptions imposed upon the regressor and error term in order to separate variation are not as needed. The LIV approach gives far less accurate and efficient results than both OLS and Gaussian copula approaches, yielding estimates with large mean bias and standard errors (e.g., $Bias^{LIV}$ = -.43; $S.E.^{LIV}$= .27); however, due to the large standard errors, the estimates are not significantly biased. This decreased accuracy and efficiency is partly due to the fact that the LIV approach is estimated through the method of maximum likelihood, which is much more sensitive to distributional misspecification.

Study 5.2: Endogenous linear regression specification. To further investigate the robustness of instrument free methods to misspecification of the error term, data was simulated from a linear regression model with an endogenous regressor and non-normal error term. The researcher specified the DGP and parameter values as follows:

[7] Although the normality assumption is not critical for estimation, it is an important assumption for inference; additionally, if we assume normality, the OLS estimator is then the best of all unbiased estimators, linear and non-linear.

$$\begin{pmatrix} x_i^* \\ \varepsilon_i^* \\ z_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix}\right),$$

$$x_i = F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\ \Phi(x_i^*)\ ) = F_{x,i}^{-1}(\ \Phi(x_i^*)\ |a, \beta\ ),$$

$$z_i = U_{z,i} = \Phi(z_i^*)\ ,$$

$$\varepsilon_i = \begin{cases} U_{\varepsilon,i} \ = \ \Phi(\varepsilon_i^*) - E(\ \Phi(\varepsilon_i^*)\ ), & Condition\ 1 \\ F_{e,i}^{-1}\left(U_{\varepsilon,i}\right) = \ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) = \ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*) \mid \nu_1, \nu_2\ ) - E\left(\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*) \mid \nu_1, \nu_2\ )\ \right), & Condition\ 2 \\ F_{e,i}^{-1}\left(U_{\varepsilon,i}\right) = \ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) = \ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*) \mid k\ ) - E\left(\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*) \mid k\ )\ \right), & Condition\ 3 \end{cases}$$

$$y_i = \beta_0 + \ \beta_1 x_i + \ \varepsilon_i = 1 + 2.5x_i + \ \varepsilon_i, \quad (56)$$

where the correlation between the regressor and error term is now $\rho_{x,e}$ = 0.5, and there is a true instrument, z_i. The error term takes on the same range of distributions as for Study 4.1. Additionally, we note that the specification of the structural error and dependence structure renders the LIV model severely misspecified for these research conditions. Means, standard errors, bias, and t_{bias} statistics are reported in Table 13.

Table 12.

Exogeneity with Misspecified Error

			OLS				LIV				Gaussian copula			
ε	**Θ**	**True Value**	**Mean**	**S.E.**	**Bias**	t_{bias}	**Mean**	**S.E.**	**Bias**	t_{bias}	**Mean**	**S.E.**	**Bias**	t_{bias}
$\sim U$	β_0	1.00	1.00	0.02	< 0.01	0.02	1.44	0.28	0.44	1.60	1.00	0.06	> -0.01	> -0.01
	β_1	2.50	2.50	0.02	> -0.01	-0.02	2.07	0.27	-0.43	-1.62	2.50	0.06	> -0.01	0.00
$\sim F(8,5)$	β_0	1.00	1.00	0.17	-0.01	-0.06	1.03	1.33	0.03	0.03	0.99	0.59	-0.01	-0.02
	β_1	2.50	2.51	0.18	0.01	0.06	2.44	1.18	-0.06	-0.05	2.51	0.60	0.01	0.02
$\sim \chi^2(4)$	β_0	1.00	1.01	0.18	0.01	0.04	0.65	1.55	-0.35	-0.23	0.97	0.59	-0.03	-0.05
	β_1	2.50	2.49	0.18	-0.01	-0.04	2.81	1.55	0.31	0.20	2.53	0.60	0.03	0.05

Table 13.

Linear Regression Endogeneity w/ Misspecified Error

			OLS				IV			
ε	**Θ**	**True Value**	**μ**	**S.E.**	**Bias**	t_{bias}	**μ**	**S.E.**	**Bias**	t_{bias}
$\sim U$	β_0	1.00	0.81	0.02	-0.19	-11.94	1.00	0.04	> -0.01	-0.01
	β_1	2.50	2.69	0.02	0.19	11.64	2.50	0.04	< 0.01	< 0.01
$\sim F(8,5)$	β_0	1.00	-0.25	0.33	-1.25	-3.76	1.03	0.34	0.03	0.08
	β_1	2.50	3.75	0.33	1.25	3.84	2.47	0.34	-0.03	-0.09
$\sim\chi^2(4)$	β_0	1.00	-0.89	0.22	-1.89	-8.76	1.02	0.40	0.02	0.05
	β_1	2.50	4.39	0.21	1.89	8.92	2.48	0.40	-0.02	-0.05

Table 13 (continued).

Linear Regression Endogeneity w/ Misspecified Error

			LIV				Gaussian copula			
ε	**Θ**	**True Value**	**μ**	**S.E.**	**Bias**	t_{bias}	**μ**	**S.E.**	**Bias**	t_{bias}
$\sim U$	β_0	1.00	1.35	0.23	0.35	1.57	1.01	0.05	0.01	0.14
	β_1	2.50	2.16	0.21	-0.34	-1.62	2.49	0.05	-0.01	-0.17
$\sim F(8,5)$	β_0	1.00	-0.62	1.61	-1.62	-1.01	-0.46	1.11	-1.46	-1.31
	β_1	2.50	4.10	1.57	1.60	1.02	3.96	1.10	1.46	1.33
$\sim\chi^2(4)$	β_0	1.00	-1.22	1.04	-2.22	-2.14	0.03	0.63	-0.97	-1.55
	β_1	2.50	4.69	1.03	2.19	2.14	3.46	0.63	0.96	1.53

Table 13 shows that OLS produces significantly biased estimates under an endogenous linear regression model with misspecification of the error. For example, when the error term takes on a Chi-square distribution with 4 degrees of freedom, the bias and t_{bias} statistics for the OLS β_1 estimate are 1.89 and 8.92, respectively. Furthermore, by looking at Tables 12 and 13 together, we see that the bias for OLS becomes considerably larger when there is endogeneity combined with an asymmetric error term. The IV method produced highly accurate, unbiased results regardless of the error distribution specified. As mentioned before, this is because the method makes use of the true instrument, which is of perfect quality and quite unlikely to be available to the researcher in a real-world setting. When the error term has a uniform distribution, the Gaussian copula produces accurate, unbiased results for the causal effect; however, when the error distribution is non-normal and asymmetric, the Gaussian copula approach produces very inaccurate estimates with large bias, albeit bias that is not statistically significantly different from zero. The non-significant *t*-statistic is in large part due to the inefficiency of the Gaussian copula estimates, as the estimates deviate substantially in value from the true parameter values. The LIV approach produces inaccurate results, having the largest bias across all non-normal error distributions. Additionally, the LIV estimate is the least efficient of all estimates. Such results suggest that the LIV model is poorly suited for situations where both the dependence structure and error term is misspecified.

Study 5.3: LIV specification. For Study 5.3, the error term was misspecified as it was in subsections 1.5.1 and 1.5.3, but now the dependence structure was specified such that it satisfied LIV model assumptions. Thus, the DGP was specified as follows:

$$\begin{pmatrix} \nu_i \\ \varepsilon_i^* \\ z_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right),$$

$$z_i = F_{z,i}^{-1}(U_{z,i}) = F_{z,i}^{-1}(\ \Phi(z_i^*)\ ) = F_{z,i}^{-1}(\ \Phi(z_i^*)\ , n, p_z\ ),$$

$$x_i = z_i\pi + \nu_i = \begin{Bmatrix} F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\ \Phi(\nu_i)\ ) = \Phi_{(6,1)}^{-1}(\ \Phi(\nu_i)\ ),\ if\ z_i = 0 \\ F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\ \Phi(\nu_i)\ ) = \Phi_{(2,1)}^{-1}(\ \Phi(\nu_i)\ ),\ if\ z_i = 1 \end{Bmatrix},$$

$$\varepsilon_i = \begin{cases} U_{\varepsilon,i} = \Phi(\varepsilon_i^*) - E(\ \Phi(\varepsilon_i^*)\ ), & Condition\ 1 \\ F_{e,i}^{-1}(U_{\varepsilon,i}) = F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) = F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*) \mid \nu_1, \nu_2\ ) - E\left(\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*) \mid \nu_1, \nu_2\ )\right), & Condition\ 2 \\ F_{e,i}^{-1}(U_{\varepsilon,i}) = F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) = F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*) \mid k\ ) - E\left(\ F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*) \mid k\ )\right), & Condition\ 3 \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5x_i + \varepsilon_i,\ (57)$$

The above specified dependence structure combined with misspecification of the error now renders the Gaussian copula model severely misspecified. Results for model performance across all four methods are presented in Table 14.

From Table 14, we see that OLS produces significantly biased estimates across all misspecified error distributions once again; this is due to the persisting endogeneity problem. The IV approach is robust to misspecification of the error given the true, ideal instrument, producing highly accurate, unbiased estimates. Of the instrument-free methods, the Gaussian copula approach yields highly inaccurate results with large bias. For the uniform and chi-square distribution, this bias is not only large but also statistically significant. These results suggest that the Gaussian copula approach is unsuitable for situations where the dependence structure differs from what the method assumes and the error is non-normal. Contrastingly, the LIV approach produces unbiased results across all misspecifications of the error. However, we notice that both the bias and standard error of the LIV estimates are much larger for the asymmetric non-normal error distributions (e.g., *F*- and Chi-square distributions). Such results suggest that the normality assumption may not be overly restrictive for this approach, but that symmetry of the error distribution matters.

Table 14.

LIV Endogeneity w/ Misspecified Error

			OLS				IV			
ε	**Θ**	**True Value**	**Mean**	**S.E.**	**Bias**	t_{bias}	**Mean**	**S.E.**	**Bias**	t_{bias}
$\sim U$	β_0	1.00	0.84	0.02	-0.16	-7.73	1.00	0.03	< 0.01	0.03
	β_1	2.50	2.54	0.01	0.04	7.81	2.50	0.01	> -0.01	-0.02
$\sim F(8,5)$	β_0	1.00	0.07	0.37	-0.93	-2.52	1.00	0.27	< 0.01	0.01
	β_1	2.50	2.73	0.09	0.23	2.55	2.50	0.07	> -0.01	-0.01
$\sim \chi^2(4)$	β_0	1.00	-0.51	0.21	-1.51	-7.28	1.00	0.26	> -0.01	-0.01
	β_1	2.50	2.88	0.05	0.38	7.49	2.50	0.06	< 0.01	0.01

Table 14 (continued).

LIV Endogeneity w/ Misspecified Error

			LIV				Gaussian copula			
ε	**Θ**	**True Value**	**Mean**	**S.E.**	**Bias**	t_{bias}	**Mean**	**S.E.**	**Bias**	t_{bias}
$\sim U$	β_0	1.00	1.09	0.10	0.09	0.94	1.62	0.09	0.62	6.59
	β_1	2.50	2.48	0.02	-0.02	-0.72	2.34	0.02	-0.16	-6.69
$\sim F(8,5)$	β_0	1.00	0.42	3.33	-0.58	-0.17	5.41	4.81	4.41	0.92
	β_1	2.50	2.62	0.68	0.12	0.18	1.39	1.19	-1.11	-0.93
$\sim \chi^2(4)$	β_0	1.00	0.20	1.14	-0.80	-0.70	7.16	1.13	6.16	5.43
	β_1	2.50	2.68	0.25	0.18	0.71	0.95	0.28	-1.55	-5.53

Study 5.4: Misspecification of LIV first stage error term. The simple LIV model assumes the existence of a discrete latent instrument and error terms that follow a joint normal distribution (Ebbes, 2004; 2009). Given the closure properties of Gaussians, it then follows that the marginal error distributions are assumed normal (Do, 2008). Therefore an interesting and worthwhile investigation is to examine the performance of the LIV model under misspecification of the first stage error term in the two-stage equation,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{(58A)}$$

$$x_i = \theta_i + \upsilon_i \quad \text{(58B)}$$

To further misspecify the LIV model, different distributions for the latent instrument can also be considered in combination with misspecification of the error term. To investigate LIV model performance under such departures from assumptions, the following DGP's were generated:

Departure 4.4.a.

$$\begin{pmatrix} \upsilon_i^* \\ \varepsilon_i \\ z_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right),$$

$$z_i = \begin{cases} 2, & if\ z_i^* < 0 \\ 6, & if\ z_i^* \geq 0 \end{cases}$$

$$\upsilon_i = \begin{cases} U_{\upsilon,i} = \Phi(\upsilon_i^*) - E(\Phi(\upsilon_i^*)), & Condition\ 1 \\ F_{\upsilon,i}^{-1}(U_{\upsilon,i}) = F_{\upsilon,i}^{-1}(\Phi(\upsilon_i^*)) = F_{\upsilon,i}^{-1}(\Phi(\upsilon_i^*) \mid g_1, g_2) - E\left(F_{\upsilon,i}^{-1}(\Phi(\upsilon_i^*) \mid g_1, g_2)\right), & Condition\ 2 \\ F_{\upsilon,i}^{-1}(U_{\upsilon,i}) = F_{\upsilon,i}^{-1}(\Phi(\upsilon_i^*)) = F_{\upsilon,i}^{-1}(\Phi(\upsilon_i^*) \mid k) - E\left(F_{\upsilon,i}^{-1}(\Phi(\upsilon_i^*) \mid k)\right), & Condition\ 3 \end{cases}$$

$$x_i = z_i + \nu_i \quad \text{(59A)}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5 x_i + \varepsilon_i \quad \text{(59B)}$$

Departure 4.4.b.

$$\begin{pmatrix} v_i \\ \varepsilon_i \\ z_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right),$$

$$z_i = \begin{cases} U_{z,i} = \Phi(z_i^*), & Condition\ 1 \\ F_{z,i}^{-1}\left(U_{z,i}\right) = F_{z,i}^{-1}(\ \Phi(z_i^*)\ ) = F_{z,i}^{-1}(\ \Phi(z_i^*) \mid v_1, v_2\ ), & Condition\ 2 \\ F_{z,i}^{-1}\left(U_{z,i}\right) = F_{z,i}^{-1}(\ \Phi(z_i^*)\ ) = F_{z,i}^{-1}(\ \Phi(z_i^*) \mid k\ ), & Condition\ 3 \end{cases}$$

$$x_i = z_i + v_i \quad \text{(60A)}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5x_i + \varepsilon_i \quad \text{(60B)}$$

Departure 4.4.c.

$$\begin{pmatrix} v_i^* \\ \varepsilon_i \\ z_i \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right),$$

$$v_i = \begin{cases} 2, & if\ v_i^* < 0 \\ 6, & if\ v_i^* \geq 0 \end{cases}$$

$$x_i = z_i + v_i \quad \text{(61A)}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5x_i + \varepsilon_i \quad \text{(61B)}$$

Tables 15, 16, and 17 summarize the results from Departures *4.4.a – c*. Given the endogeneity problem across the specified conditions, OLS remains biased, producing estimates that are tightly distributed around an erroneous value. This is consistent with all previous simulation results and thus will not be discussed in more detail. Provided with the true, ideal instrument, we see that the IV approach continues to produce highly accurate, unbiased estimates of the true parameter value regardless of the DGP specification. Examining the performance of the instrument-free methods, we see that the Gaussian copula approach produces inaccurate estimates with large bias when the first-stage error has a uniform and *F* distribution; however, when the first-stage error has a chi-square distribution, the approach seems to gain accuracy, producing a parameter estimate with small bias. Furthermore, we see that regardless of the distribution specified for the latent instrument, the Gaussian copula approach yields inaccurate results with large bias. And when data is simulated from a LIV model with both misspecified first-stage error term and latent instrument, the Gaussian copula approach produces estimates that are very inaccurate and significantly biased away from true parameter values.

The LIV approach seems to be somewhat robust to misspecification of the first-stage error term. Across all three first-stage error specifications, the LIV method produced statistically unbiased results; however, as seen with misspecification of the structural error term, the LIV estimates have larger bias and standard errors when the first-stage error term takes on asymmetric non-normal distributions. Surprisingly, the LIV approach produces accurate, unbiased results despite misspecification of the latent instrument. Such results indicate that the latent instrument assumption of the simple LIV model may not be overly restrictive, as model performance appears to be relatively unaffected under violations of this assumption. When both the first-stage error and latent instrument are misspecified, however, the LIV approach produces significantly biased results as also seen with OLS and Gaussian copula approaches.

Table 15.

Misspecified first-stage error

			OLS				IV			
υ	**Θ**	**True Value**	**Mean**	**S.E.**	**Bias**	t_{bias}	**Mean**	**S.E.**	**Bias**	t_{bias}
$\sim U$	β_0	1.00	0.80	0.10	-0.20	-2.03	1.00	0.10	> -0.01	-0.05
	β_1	2.50	2.55	0.02	0.05	2.23	2.50	0.02	< 0.01	0.03
$\sim F(8,5)$	β_0	1.00	0.51	0.11	-0.49	-4.52	1.00	0.09	< 0.01	0.05
	β_1	2.50	2.62	0.02	0.12	5.01	2.50	0.02	> -0.01	-0.04
$\sim \chi^2(4)$	β_0	1.00	0.37	0.06	-0.63	-10.39	1.00	0.10	< 0.01	0.03
	β_1	2.50	2.66	0.01	0.16	15.75	2.50	0.02	> -0.01	-0.02

Table 15 (continued).

Misspecified first-stage error

			LIV				Gaussian copula			
υ	**Θ**	**True Value**	**Mean**	**S.E.**	**Bias**	t_{bias}	**Mean**	**S.E.**	**Bias**	t_{bias}
$\sim U$	β_0	1.00	1.22	0.30	0.22	0.74	2.93	0.18	1.93	10.98
	β_1	2.50	2.46	0.07	-0.04	-0.55	2.01	0.04	-0.49	-11.44
$\sim F(8,5)$	β_0	1.00	0.69	0.76	-0.31	-0.41	1.64	0.48	0.64	1.35
	β_1	2.50	2.58	0.16	0.08	0.47	2.34	0.12	-0.16	-1.35
$\sim \chi^2(4)$	β_0	1.00	0.26	0.46	-0.74	-1.61	0.87	0.29	-0.13	-0.44
	β_1	2.50	2.69	0.12	0.19	1.59	2.53	0.07	0.03	0.42

Table 16.

Misspecified latent instrument

			OLS				IV			
$\tilde{z}$	Θ	True Value	Mean	S.E.	Bias	t_{bias}	Mean	S.E.	Bias	t_{bias}
$\sim U$	β_0	1.00	0.68	0.04	-0.32	-8.84	1.01	0.10	0.01	0.09
	β_1	2.50	3.15	0.03	0.64	20.26	2.49	0.16	-0.01	-0.09
$\sim F(8,5)$	β_0	1.00	0.80	0.08	-0.20	-2.39	1.00	0.06	> -0.01	-0.01
	β_1	2.50	2.62	0.05	0.12	2.44	2.50	0.02	< 0.01	0.06
$\sim \chi^2(4)$	β_0	1.00	0.69	0.07	-0.31	-4.21	1.00	0.08	> -0.01	-0.04
	β_1	2.50	2.58	0.02	0.08	4.87	2.50	0.02	< 0.01	0.05

Table 16 (continued).

Misspecified latent instrument

			LIV				Gaussian copula			
$\tilde{z}$	Θ	True Value	Mean	S.E.	Bias	t_{bias}	Mean	S.E.	Bias	t_{bias}
$\sim U$	β_0	1.00	0.99	0.09	-0.01	-0.15	0.67	0.15	-0.33	-2.17
	β_1	2.50	2.53	0.17	0.03	0.18	3.15	0.30	0.65	2.16
$\sim F(8,5)$	β_0	1.00	0.96	0.17	-0.04	-0.24	1.28	0.15	0.28	1.87
	β_1	2.50	2.52	0.09	0.02	0.19	2.32	0.10	-0.18	-1.86
$\sim \chi^2(4)$	β_0	1.00	0.83	0.25	-0.17	-0.68	1.68	0.27	0.68	2.52
	β_1	2.50	2.54	0.06	0.04	0.67	2.33	0.07	-0.17	-2.57

Table 17.

Misspecified latent instrument and first-stage error

		OLS				IV				LIV				Gaussian copula			
Θ	**True Value**	**Mean**	**S.E.**	**Bias**	t_{bias}	**Mean**	**S.E.**	**Bias**	t_{bias}	**Mean**	**S.E.**	**Bias**	t_{bias}	**Mean**	**S.E.**	**Bias**	t_{bias}
β_0	1.00	0.10	0.08	-0.90	-11.88	1.01	0.19	0.01	0.03	0.04	0.27	-0.96	-3.60	-0.98	0.28	-1.98	-7.02
β_1	2.50	2.72	0.02	0.22	13.10	2.50	0.05	> -0.01	-0.04	2.74	0.06	0.24	4.03	3.00	0.07	0.50	7.09

Study 6: Impact of Sample Size and Instrument Quality

Study 6 investigates the impact of sample size on the performance of instrumental variable regression compared with that of instrument-free methods. The comparisons are described as sub-studies of Study 6. To perform these sub-studies, the researcher simulated endogeneity from three DGP's and varied the sample size, *N*, for each DGP as follows: *50*, *100, 250*, *500*, *1000*, *2500*, *and 5000*. Such values for *N* reflect sample sizes ranging from very small to considerably large. The DGP's and true parameter values were specified as follows:

DGP 6.a: Linear Regression Model

$$\begin{pmatrix} x_i^* \\ \varepsilon_i^* \\ z_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{x,e} & \rho_{x,z} \\ \rho_{x,e} & 1 & \rho_{z,e} \\ \rho_{x,z} & \rho_{z,e} & 1 \end{bmatrix} \right),$$

$$x_i = F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\ \Phi(x_i^*)\ ) = F_{x,i}^{-1}(\ \Phi(x_i^*)\ |2,2\ ) - E(\ F_{x,i}^{-1}(\ \Phi(x_i^*)\ |2,2\ )\ )\ ,$$

$$\varepsilon_i = F_{e,i}^{-1}(U_{\varepsilon,i}) = F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) = \Phi^{-1}(\ \Phi(\varepsilon_i^*)\ ),$$

$$z_i = U_{z,i} = \Phi(z_i^*)\ ,$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5x_i + \varepsilon_i, \quad (62)$$

DGP 6.b: Latent Instrument is Observed Instrument LIV Model

$$\begin{pmatrix} \nu_i \\ \varepsilon_i^* \\ \tilde{z}_i^* \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right),$$

$$\tilde{z} = F_{\tilde{z},i}^{-1}(U_{\tilde{z},i}) = F_{\tilde{z},i}^{-1}(\ \Phi(\tilde{z}_i^*)\ ) = F_{\tilde{z},i}^{-1}(\ \Phi(\tilde{z}_i^*)\ , 1,\ 0.5),$$

$$x_i = \tilde{z}\pi + \nu_i = \begin{cases} F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\ \Phi(\nu_i)\ ) = \Phi_{(6,1)}^{-1}(\ \Phi(\nu_i)\ ),\ if\ \tilde{z} = 1 \\ F_{x,i}^{-1}(U_{x,i}) = F_{x,i}^{-1}(\ \Phi(\nu_i)\ ) = \Phi_{(2,1)}^{-1}(\ \Phi(\nu_i)\ ),\ if\ \tilde{z} = 0 \end{cases}, \rightarrow x_i = x_i - E(\ x_i\ ),$$

$$\varepsilon_i = F_{e,i}^{-1}(U_{\varepsilon,i}) = F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) = \Phi^{-1}(\ \Phi(\varepsilon_i^*)\ ),$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1 + 2.5x_i + \varepsilon_i, \quad (63)$$

DGP 6.c: LIV Model with Observed Instrument

$$\begin{pmatrix} \upsilon_i \\ \varepsilon_i^* \\ \tilde{z}_i^* \\ z_i \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{\upsilon,e} & 0 & 0 \\ \rho_{\upsilon,e} & 1 & 0 & \rho_{z,e} \\ 0 & 0 & 1 & \rho_{\tilde{z},z} \\ 0 & \rho_{z,e} & \rho_{\tilde{z},z} & 1 \end{bmatrix}\right),$$

$$\tilde{z}_i = F_{\tilde{z},i}^{-1}(U_{\tilde{z},i}) = F_{\tilde{z},i}^{-1}(\ \Phi(\tilde{z}_i^*)\ ) = F_{\tilde{z},i}^{-1}(\ \Phi(\tilde{z}_i^*)\ , 1,\ 0.5),$$

$$x_i = \tilde{z}_i\pi + \upsilon_i \ = \begin{Bmatrix} F_{x,i}^{-1}(U_{x,i}) = \ F_{x,i}^{-1}(\ \Phi(\upsilon_i)\ ) = \ \Phi_{(6,1)}^{-1}(\ \Phi(\upsilon_i)\ ),\ if\ \tilde{z}_i = 1 \\ F_{x,i}^{-1}(U_{x,i}) = \ F_{x,i}^{-1}(\ \Phi(\upsilon_i)\ ) = \ \Phi_{(2,1)}^{-1}(\ \Phi(\upsilon_i)\ ),\ if\ \tilde{z}_i = 0 \end{Bmatrix} \rightarrow x_i = x_i - E(\ x_i\ ),$$

$$\varepsilon_i = F_{e,i}^{-1}(U_{\varepsilon,i}) = F_{e,i}^{-1}(\ \Phi(\varepsilon_i^*)\ ) = \Phi^{-1}(\ \Phi(\varepsilon_i^*)\ ),$$

$$y_i = \beta_0 + \ \beta_1 x_i + \ \varepsilon_i = 1 + 2.5x_i + \ \varepsilon_i,\ (64)$$

Five-hundred data sets were independently generated for each sample size *N*. Additionally, the degree of endogeneity, relevance and validity of the observed instrument were varied in order to compare performance of the instrument-free methods with the instrumental variable method across a range of different scenarios in combination with sample size. For the linear regression specification (DGP 6.a), endogeneity, denoted $\rho_{x,e}$, was set equal to 0.1 and 0.5, representing minor and severe degrees of endogeneity, respectively. Likewise, instrument relevance, $\rho_{x,z}$, was set equal to 0.1 and 0.5, representing a weak to strong instrument, respectively. Instrument validity, $\rho_{z,e}$, was set equal to 0.0, 0.1, and 0.5, representing a valid instrument; an invalid, slightly endogenous instrument; and an invalid, severely endogenous instrument.

For the LIV specification DGP 6.1b, endogeneity was generated from a simple LIV model with a discrete, relevant, and exogenous latent instrument. Furthermore, the latent instrument comprising the exogenous part of x_i was also used as the observed instrument in the instrumental analyses performed under DGP 6.b. This represents the rare, yet ideal scenario where the latent instrument manifests itself as the observed instrument, i.e., $\tilde{z}_i = z_i$, and therefore the perfect, true instrument is available to the researcher.

For the LIV specification DGP 6.1c, the exogenous variation in x_i is also captured by the discrete latent instrument, $\tilde{z}_i$; however, this latent variable now remains unobserved and instead an observed instrument, z_i, is available to the researcher that varies in respect to validity and relevance[8]. This represents the arguably more common research scenario where the researcher searches for and identifies an available instrument that is not the true instrument but serves as a proxy and is of variable quality. For this DGP, instrument relevance, $\rho_{\tilde{z},z}$, was set equal to 0.1 and 0.7, representing a weak and strong instrument, respectively. The values for instrument relevance, $\rho_{\tilde{z},z}$, differ from those in the linear regression specification because both the non-linear transformation of $\tilde{z}_i^*$ and the structural specification of x_i attenuates the specified correlation[9]; level of endogeneity, denoted $\rho_{v,e}$, was set equal to 0.2 and 0.8, representing minor and severe degrees of endogeneity, respectively. These values also differ from $\rho_{x,e}$ in the linear regression specification because the endogenous regressor, x_i, for the LIV model will be correlated with the structural error by roughly $\frac{1}{2} * \rho_{v,e}$, given the specification. Lastly, instrument validity, $\rho_{z,e}$, was specified the same as it was for the linear regression specification.

[8] Varying the parameters in the correlation matrix resulting from this specification led to 12 matrices, of which 1 was non-positive definite, i.e., it would not be a population correlation matrix; this was handled by computing the nearest positive definite matrix and using a correlation matrix based on this result.
[9] This attenuation happens especially for higher values of the correlation

For all DGPs, we note that the endogenous regressor, x_i, was mean-centered in order to allow for unbiased estimation of the intercept across all methods; as a result, the researcher omits results for the intercept for this section and only focuses on information regarding the regression coefficient, β_1. This was done in order to limit the amount of output presented, as given the number of parameters varied for this section, the output was considerable. The statistic of central focus for all following sections is Mean Squared Error (MSE), which is a measure of how far an estimate is from the true parameter value on average (Boef, Dekkers, Vandenbroucke, & le Cessie, 2014). In mathematical terms, this statistic is the average of the squared deviations of an estimate from the true parameter value, and is written as

$$MSE = E[(\hat{\beta} - \beta)^2] = bias^2 + variance \qquad (65)$$

The MSE was chosen as a performance measure because it simultaneously takes into account both the bias and variance of an estimate. A lower MSE value means that an estimate is, on average, closer to the true parameter value and thereby indicates better estimation performance.

Study 6.1: High quality instruments. The performance of OLS, IV, and the Gaussian copula methods were first compared with data generated from DGP 6.a, representing the condition where there is a highly relevant, valid instrument, i.e., a perfect instrument. Table 18 provides MSE statistics for the three methods. We see that when endogeneity is minor, i.e., $\rho_{x,e} = 0.10$, OLS produces the estimate that is on average closest to the true parameter value for $N \leq 500$. This is largely due to the smaller variance of the OLS estimate at smaller sample sizes; additionally, the minor endogeneity condition biases OLS parameter estimates only slightly when compared with more severe endogeneity conditions. Consistent with the extant literature on instrumental variables, results show that IV regression outperforms OLS at larger sample sizes, $N > 500$, even under the most ideal conditions where a perfect instrument is available. Moreover, we see that the Gaussian copula regression approach also requires larger sample sizes, outperforming OLS at $N > 1000$ and equaling the performance of IV regression with a perfect instrument at $N \geq 5000$. As expected, however, both the instrumental variable and Gaussian copula regression approaches outperform OLS at much smaller sample sizes when endogeneity becomes more severe, i.e., $\rho_{x,e} = 0.50$; specifically, IV with a perfect instrument outperforms OLS at an $N = 50$ and the Gaussian copula approach outperforms OLS at $N = 100$. Thus, we see that both the degree of unmeasured confounding and sample size matter when determining performance of instrumental variable and the Gaussian copula approaches relative to OLS. Additionally, we note that the Gaussian copula instrument-free approach never outperforms IV regression in terms of MSE when a perfect instrument is available to the researcher.

Table 18.

*Strong, Valid Instrument under Linear Regression model**

		Mean Squared Error		
N	$\boldsymbol{\rho_{x,e}}$	**OLS**	**IV**	**Gaussian copula**
50	0.1	0.06	0.25	0.55
	0.5	0.51	0.32	0.66
100	0.1	0.04	0.11	0.29
	0.5	0.49	0.12	0.28
250	0.1	0.03	0.04	0.10
	0.5	0.47	0.04	0.10
500	0.1	0.02	0.02	0.04
	0.5	0.46	0.02	0.04
1,000	0.1	0.02	0.01	0.02
	0.5	0.45	0.01	0.02
2,500	0.1	0.02	< .01	0.01
	0.5	0.45	< .01	0.01
5,000	0.1	0.02	< .01	< .01
	0.5	0.45	< .01	< .01

***True parameter value is 2.5**

The researcher next compared the performance of the LIV instrument-free method with that of OLS and instrumental variable regression with a perfect instrument. Table 19 gives MSE statistics for data generated from DGP 6.b, representing the scenario where the latent instrument manifests itself as the observed instrument, and thus the researcher once again has a perfect instrument available for use. Moreover, there is moderate to severe endogeneity for this condition (i.e., $\rho_{\varepsilon^*,\upsilon} = 0.7$).

Table 19.

*Latent instrument is observed instrument**

	Mean Squared Error		
N	**OLS**	**IV**	**LIV**
50	0.022	0.006	0.006
100	0.020	0.003	0.002
250	0.020	0.001	0.001
500	0.020	0.001	0.001
1,000	0.019	< .001	< .001
2,500	0.020	< .001	< .001
5,000	0.020	< .001	< .001

***True parameter value is 2.5**

Similar to the results comparing OLS, IV and the Gaussian copula methods, we see that both IV regression and the instrument-free LIV method outperform OLS at small sample sizes, N = 50, when endogeneity is severe. For example, the MSE for IV and LIV is .006, whereas for OLS the MSE is 0.022. More interestingly, we see that the performance of the LIV instrument-free method equals that of the IV method across roughly all sample sizes. This suggests that the LIV method produces estimates that are, on average, as close to the true parameter value as estimates from an IV regression using a perfect instrument. This finding is conditional on the assumptions of the LIV model being satisfied[10].

Subsequently, the performance of the LIV instrument-free method was compared with that of OLS and IV regression approaches under varying degrees of endogeneity. A strong, valid instrument remains available for use with IV regression; however, for this condition, the observed instrument is no longer the latent instrument, but is instead a

[10] Specifically, there exists a discrete latent instrument that is uncorrelated with the error term and there is additive separability in the endogenous regressor.

proxy for it, albeit a high quality one. Table 20 provides MSE statistics across all three methods as a comparative measure.

Table 20.

*High quality observed instrument under LIV model**

		Mean Squared Error		
N	$\boldsymbol{\rho_{v,e}}$	**OLS**	**IV**	**LIV**
50	0.2	0.005	0.019	0.006
	0.8	0.028	0.022	0.005
100	0.2	0.003	0.009	0.003
	0.8	0.027	0.008	0.002
250	0.2	0.002	0.003	0.001
	0.8	0.026	0.004	0.001
500	0.2	0.002	0.002	0.001
	0.8	0.026	0.002	0.001
1,000	0.2	0.002	0.001	< .001
	0.8	0.026	0.001	< .001
2,500	0.2	0.002	< .001	< .001
	0.8	0.025	< .001	< .001
5,000	0.2	0.002	< .001	< .001
	0.8	0.026	< .001	< .001

***True parameter value is 2.5**

Consistent with the researcher's previous findings, we see that IV regression outperforms OLS at larger sample sizes, $N > 500$, when there exists minor endogeneity. Interestingly, however, is that the instrument-free LIV method outperforms OLS at $N > 100$ for the minor endogeneity condition (i.e., $\rho_{v,e}$ = 0.2). Moreover, the LIV approach outperforms the IV regression approach across all sample sizes. This is due to the IV approach no longer using the true latent instrument but instead using an observed instrument. Although the observed instrument is of very high quality, there is an efficiency cost for using an observed variable as a proxy for the true latent instrument, and thus the method underperforms the LIV method. These findings, in conjunction with the findings from Table 19, suggest that the LIV method is comparable to an IV method with a true, perfect instrument. Findings from the severe endogeneity condition, $\rho_{\upsilon,e}$ = 0.8, are similar as those from $\rho_{\upsilon,e}$ = 0.2; however, we now see that the IV regression approach outperforms OLS at much smaller sample sizes, $N = 50$, as was found with earlier comparisons. Additionally, the instrument-free LIV method outperforms both OLS and IV regression with an observed high quality instrument across all sample sizes when severe endogeneity exists.

Study 6.2: Weak instruments. The performance of OLS, IV with a weak, valid instrument, and instrument-free methods was compared across various sample sizes. To specifically compare OLS, IV with a weak instrument, and the Gaussian copula method, data was generated from DGP 6.a with $\rho_{x,z}$ = 0.10. This represents the likely scenario where the researcher, in an attempt to satisfy the exogeneity requirement, identifies an instrument that is only weakly correlated with the endogenous regressor. Table 21 provides MSE statistics across the three methods.

Table 21.

*Weak, valid instrument under Linear Regression model**

		Mean Squared Error		
N	$\rho_{x,e}$	**OLS**	**IV**	**Gaussian copula**
50	0.1	0.06	3991.10	0.62
	0.5	0.51	489.08	0.64
100	0.1	0.04	3145.71	0.28
	0.5	0.5	226.88	0.28
250	0.1	0.02	345.28	0.09
	0.5	0.46	177.48	0.08
500	0.1	0.02	538.04	0.05
	0.5	0.45	61.04	0.04
1,000	0.1	0.02	1.3	0.02
	0.5	0.45	0.51	0.02
2,500	0.1	0.02	0.11	0.01
	0.5	0.45	0.13	0.01
5,000	0.1	0.02	0.05	< .01
	0.5	0.45	0.06	< .01

***True parameter value is 2.5**

For the minor endogeneity condition, $\rho_{x,e}$= 0.10, we see that once again the instrument-free Gaussian copula method requires larger sample sizes, outperforming OLS at $N >$ 1000, which is consistent with the researcher's previous findings. Notably, however, the Gaussian copula approach now outperforms the IV method across all sample sizes. This is interesting because the instrument used for the IV method is still a valid instrument. Given the weak instrument, the IV method also underperforms the OLS method across all sample sizes as well.

For the severe endogeneity condition, $\rho_{x,e}$= 0.50, we note that the IV method requires very large sample sizes, outperforming OLS at $N \geq$ 2,500. However, the IV regression approach still never outperforms the Gaussian copula method when the instrument is weak, regardless of the degree of endogeneity specified. Furthermore, given the severe degree of endogeneity present, the Gaussian copula approach now outperforms

OLS at much smaller sample sizes, producing estimates that are on average closer to the true parameter value for $N \geq 100$. Such result suggest that using a weak instrument under the IV approach still produces consistent estimates, however, very large sample sizes are required before the method is of value. In such instances, it is far more beneficial to use the instrument-free Gaussian copula approach, which produces much more accurate results at far smaller sample sizes.

The researcher next generated data from DGP 6.c with $\rho_{\tilde{z},z}$ = 0.2, representing a weak observed instrument. Table 22 provides MSE statistics for OLS, IV, and LIV models.

Table 22.

*Weak, valid instrument under LIV model**

		Mean Squared Error		
N	$\boldsymbol{\rho_{v,e}}$	**OLS**	**IV**	**LIV**
50	0.2	0.006	14.691	0.007
	0.8	0.028	301.462	0.005
100	0.2	0.004	7.687	0.003
	0.8	0.027	12.462	0.002
250	0.2	0.002	93.141	0.001
	0.8	0.026	29.1	0.001
500	0.2	0.002	1.246	0.001
	0.8	0.026	4.138	0.001
1,000	0.2	0.002	0.323	< .001
	0.8	0.025	0.485	< .001
2,500	0.2	0.002	0.021	< .001
	0.8	0.026	0.024	< .001
5,000	0.2	0.002	0.009	< .001
	0.8	0.026	0.009	< .001

***True parameter value is 2.5**

Once again, we see that the IV model performs the worst under the weak endogeneity condition of $\rho_{v,e}$ = 0.2, underperforming both OLS and LIV approaches across all sample sizes. The instrument-free LIV approach outperforms both OLS and IV at $N \geq 100$, producing estimates that are on average closest to the true parameter value for these sample sizes. For the major endogeneity condition of $\rho_{v,e} = 0.8$, results for the IV regression approach are consistent with earlier results, with the IV regression requiring very large sample sizes, $N \geq 2{,}500$, before outperforming OLS. Additionally, IV with a weak instrument never outperforms the LIV method, regardless of the degree of endogeneity specified. Such results suggest that the LIV approach is a suitable alternative to IV regression methods, especially when the instrument is valid but weak and the sample size is small.

Study 6.3: Strong, invalid instruments. Instrument quality in combination with sample size was once again varied, and the researcher generated data from DGP 6.a with $\rho_{x,z}$ = 0.5 and $\rho_{z,e}$ = 0.10, 0.50. Such values represent a strong but invalid instrument and the likely scenario of the researcher using an endogenous instrument due to trying to satisfy the relevance requirement. Table 23 provides the MSE statistics for this condition.

For the minor endogeneity condition, $\rho_{x,e}$ = 0.10, results in Table 23 once again show that the Gaussian copula method outperforms OLS at $N \geq 1000$. In regards to instrument quality, we see that when minor endogeneity is combined with instrument invalidity, IV never outperforms OLS, yielding estimates farther from the true parameter value on average. This is true regardless of the degree of instrument validity specified, i.e., $\rho_{z,e} = 0.10$ and 0.50. Comparing IV to the instrument-free Gaussian copula method, we see that instrument validity matters more. When there is minor endogeneity and the instrument is only slightly invalid, the Gaussian copula method outperforms IV regression at $N \geq 250$; however, when the instrument becomes strongly invalid, the Gaussian copula method outperforms IV regression across all sample sizes.

Table 23.

*Strong, invalid instrument under Linear Regression model**

Mean Squared Error					
N	**$\rho_{z,e}$**	**$\rho_{x,e}$**	**OLS**	**IV**	**Gaussian copula**
50	0.1	0.1	0.06	0.44	0.60
		0.5	0.51	0.33	0.76
	0.5	0.1	0.07	7.37	0.62
		0.5	0.51	2.80	0.69
100	0.1	0.1	0.04	0.22	0.31
		0.5	0.48	0.19	0.29
	0.5	0.1	0.04	2.80	0.27
		0.5	0.48	2.42	0.28
250	0.1	0.1	0.03	0.14	0.10
		0.5	0.46	0.11	0.10
	0.5	0.1	0.03	2.38	0.10
		0.5	0.47	2.34	0.09
500	0.1	0.1	0.02	0.11	0.05
		0.5	0.46	0.10	0.04
	0.5	0.1	0.02	2.29	0.04
		0.5	0.46	2.29	0.04
1,000	0.1	0.1	0.02	0.10	0.02
		0.5	0.45	0.10	0.02
	0.5	0.1	0.02	2.31	0.02
		0.5	0.45	2.24	0.02
2,500	0.1	0.1	0.02	0.09	0.01
		0.5	0.45	0.09	0.01
	0.5	0.1	0.02	2.24	0.01
		0.5	0.45	2.25	0.01
5,000	0.1	0.1	0.02	0.09	< .01
		0.5	0.45	0.09	< .01
	0.5	0.1	0.02	2.25	< .01
		0.5	0.45	2.25	< .01

***True parameter value is 2.5**

When degree of endogeneity is increased, $\rho_{x,e}$ = 0.50, we notice slightly different results. Given severe endogeneity and only minor instrument invalidity, i.e., $\rho_{z,e} = 0.10$, we see that IV regression now outperforms OLS across all sample sizes. This is due to the large bias in the OLS estimates resulting from severe endogeneity. As before, the

Gaussian copula method outperforms the IV approach at $N \geq 250$. However, when instrument invalidity is increased, $\rho_{z,e}$ = 0.50, IV regression now underperforms OLS across all sample sizes, indicating that the bias in the IV estimates that results from the invalidity of the observed instrument is worse than the bias in the OLS estimates resulting from severe endogeneity. Additionally, when instrument invalidity is severe, the instrument-free Gaussian copula method outperforms IV regression across all sample sizes. Given severe endogeneity, the Gaussian copula method once again outperforms OLS for $N \geq 100$. Overall, such results suggest that the Gaussian copula approach is a suitable alternative to both OLS and IV regression for even small to moderate sample sizes when there is a high degree of endogeneity and a highly valid instrument is unavailable.

Instrument validity in combination with sample size was varied under the LIV model in order to compare OLS, IV, and LIV approaches. The researcher now generated data from DGP 6.c with $\rho_{\tilde{z},z}$ = 0.7 and $\rho_{z,e}$ = 0.10 and 0.50. Table 24 provides MSE statistics for OLS, IV, and LIV under given specifications. For the minor endogeneity condition, $\rho_{\upsilon,e}$ = 0.20, we note that the LIV method outperforms OLS at $N \geq 250$. Interestingly, when minor endogeneity is combined with instrument invalidity, we see that IV regression underperforms both the OLS and LIV approaches across all sample sizes, yielding estimates that are on average farthest from the true parameter value. This is true regardless of the level of instrument invalidity specified, i.e., $\rho_{z,e}$ = 0.10 and 0.50. This is largely due to the fact that the IV regression makes use of not only an invalid instrument, thus biasing its estimates, but also an observed proxy instead of the true, latent instrument, thereby incurring an efficiency cost.

Table 24.

*Strong, invalid instrument under LIV model**

Mean Squared Error					
N	$\boldsymbol{\rho_{z,e}}$	$\boldsymbol{\rho_{v,e}}$	**OLS**	**IV**	**LIV**
50	0.1	0.2	0.006	0.028	0.006
		0.8	0.028	0.023	0.005
	0.5	0.2	0.005	0.263	0.006
		0.8	0.027	0.210	0.005
100	0.1	0.2	0.003	0.018	0.003
		0.8	0.026	0.014	0.002
	0.5	0.2	0.004	0.211	0.003
		0.8	0.027	0.196	0.002
250	0.1	0.2	0.002	0.011	0.001
		0.8	0.026	0.011	0.001
	0.5	0.2	0.002	0.207	0.001
		0.8	0.026	0.191	0.001
500	0.1	0.2	0.002	0.009	0.001
		0.8	0.026	0.009	< .001
	0.5	0.2	0.002	0.207	0.001
		0.8	0.026	0.190	0.001
1,000	0.1	0.2	0.002	0.009	< .001
		0.8	0.026	0.009	< .001
	0.5	0.2	0.002	0.203	< .001
		0.8	0.026	0.185	< .001
2,500	0.1	0.2	0.002	0.008	< .001
		0.8	0.026	0.008	< .001
	0.5	0.2	0.002	0.201	< .001
		0.8	0.026	0.184	< .001
5,000	0.1	0.2	0.002	0.008	< .001
		0.8	0.026	0.008	< .001
	0.5	0.2	0.002	0.200	< .001
		0.8	0.026	0.184	< .001

***True parameter value is 2.5**

When the degree of endogeneity is increased, $\rho_{v,e}$ = 0.80, we once again see that instrument validity matters more. We see that when severe endogeneity is combined with minor instrument invalidity, IV regression outperforms OLS across all sample sizes; however, when the invalidity of the instrument is increased, $\rho_{z,e}$ = 0.50, OLS now outperforms IV regression across all sample sizes, once again indicating that the bias in the IV estimates resulting from the invalidity of the observed instrument is worse than the bias in the OLS estimates resulting from severe endogeneity. Importantly, we note that regardless of the invalidity of the instrument specified, the IV regression approach never outperforms the LIV method. Such results suggest that the LIV approach is a suitable alternative to both OLS and IV for even small to moderate sample sizes when endogeneity is present and a valid instrument is unavailable.

Study 6.4: Weak, invalid instruments. The researcher investigated the impact of sample size on OLS, IV with weak, invalid instruments, and instrument-free methods. In comparing OLS, IV, and the Gaussian copula approach, data was generated from DGP 6.a with $\rho_{x,z}$ = 0.1 and $\rho_{z,e}$ = 0.10, 0.50. This represents the scenario where the researcher has identified an overall poor quality instrument, as it is both weak and invalid. Table 25 provides MSE statistics across sample sizes for this scenario.

From Table 25, we can see that the IV approach underperforms both OLS and Gaussian copula approaches across all sample sizes, regardless of degree of endogeneity specified or level of instrument invalidity specified. For example, for a sample size of 5,000, instrument endogeneity of .10, and regressor endogeneity of .50, we see that MSE is .45, 2.38, and < .01 for OLS, IV, and Gaussian copula methods, respectively. Such results indicate that the combination of low instrument relevance and instrument invalidity render the IV model highly inaccurate, regardless of how large the sample becomes. We see that for the minor endogeneity condition, $\rho_{x,e}$ = 0.10, the Gaussian copula approach outperforms OLS at $N > 1000$; when endogeneity is severe, $\rho_{x,e}$ = 0.50, the Gaussian copula approach outperforms OLS at $N \geq 100$. To compare the OLS, IV, and LIV models, data was generated from DGP 6.c with $\rho_{\tilde{z},z}$ = 0.10 and $\rho_{z,e}$ = 0.10 and 0.50. This represents the scenario where a researcher does not have the true, unobserved instrument available to her, and instead relies on a poor quality instrument, as it is both weak and invalid.

Results from Table 26 show that the IV regression approach underperforms both OLS and LIV approaches across all sample sizes, regardless of degree of endogeneity or level of instrument invalidity specified. These results confirm that the combination of low instrument relevance and instrument invalidity render the IV model highly inaccurate, regardless of how large the sample becomes. We see that for the minor endogeneity condition, $\rho_{\upsilon,e}$ = 0.20, the LIV approach consistently outperforms OLS at $N \geq 250$. When endogeneity is increased to $\rho_{\upsilon,e}$ = 0.80, the LIV approach outperforms OLS across all sample sizes, as seen in previous results.

When viewed holistically, results from Study 5 highlight the importance of considering a variety of factors when performing causal inference analyses. While the instrumental variable approach is useful for addressing endogeneity bias, its utility is largely dependent on the quality of the instrument. Furthermore, as simulation results have shown, even when an instrument has desirable properties, namely relevance and validity, it may produce estimates farther on average from the true parameter value than OLS if the sample size is insufficient. The utility of alternative instrument-free approaches was also investigated and simulation results have shown that such methods can successfully circumvent the need for identifying an observed instrument with certain desirable properties, as needed for IV analyses. Moreover, it was shown that instrument-free methods often perform better than instrumental variable regression when the instrument is either weak, invalid or a combination of the two. However, such findings are contingent upon the assumptions of these models being met, and sample size remains an important factor to consider. For the Gaussian copula method, larger sample sizes are often needed, especially if there is only minor endogeneity. For the LIV approach, simulation results suggest that only moderate sample sizes are needed when there is minor endogeneity. However, when the degree of endogeneity increases, both instrument-free methods outperform OLS at much smaller sample sizes, despite the smaller variance of the latter. Furthermore, when the instrument is anything but perfect, both instrument-free approaches may outperform IV regression, even at very small sample sizes.

Table 25.

*Weak, invalid instrument under Linear Regression model**

			Mean Squared Error		
N	$\rho_{z,e}$	$\rho_{x,e}$	**OLS**	**IV**	**Gaussian copula**
50	0.1	0.1	0.06	2801.44	0.63
		0.5	0.49	300.84	0.64
	0.5	0.1	0.06	11237.83	0.59
		0.5	0.51	660539.6	0.69
100	0.1	0.1	0.05	232.03	0.29
		0.5	0.49	157.51	0.31
	0.5	0.1	0.04	26860.66	0.29
		0.5	0.47	2426.69	0.31
250	0.1	0.1	0.03	55.4	0.09
		0.5	0.47	154.22	0.10
	0.5	0.1	0.03	394232.2	0.10
		0.5	0.47	335707.3	0.09
500	0.1	0.1	0.02	53.14	0.05
		0.5	0.45	25.47	0.03
	0.5	0.1	0.02	67141.92	0.05
		0.5	0.45	3300.93	0.05
1,000	0.1	0.1	0.02	383.32	0.02
		0.5	0.45	7.57	0.02
	0.5	0.1	0.02	133.46	0.02
		0.5	0.46	325.63	0.02
2,500	0.1	0.1	0.02	2.67	0.01
		0.5	0.45	2.54	0.01
	0.5	0.1	0.02	65.56	0.01
		0.5	0.45	66.11	0.01
5,000	0.1	0.1	0.02	2.47	< .01
		0.5	0.45	2.38	< .01
	0.5	0.1	0.02	59.79	< .01
		0.5	0.45	59	< .01

***True parameter value is 2.5**

Table 26.

*Weak, invalid instrument under LIV model**

Mean Squared Error					
N	$\rho_{z,e}$	$\rho_{v,e}$	**OLS**	**IV**	**LIV**
50	0.1	0.2	0.005	6133.535	0.006
		0.8	0.027	42.473	0.006
	0.5	0.2	0.006	3095.335	0.006
		0.8	0.027	5379.293	0.006
100	0.1	0.2	0.004	38.088	0.003
		0.8	0.025	10.114	0.002
	0.5	0.2	0.003	4086.72	0.003
		0.8	0.026	945.065	0.002
250	0.1	0.2	0.002	51.509	0.001
		0.8	0.026	1146.919	0.001
	0.5	0.2	0.002	13387.95	0.001
		0.8	0.026	655.788	0.001
500	0.1	0.2	0.002	97.469	0.001
		0.8	0.026	63.740	< .001
	0.5	0.2	0.002	3164.579	0.001
		0.8	0.026	863.613	< .001
1,000	0.1	0.2	0.002	9.746	< .001
		0.8	0.026	5.604	< .001
	0.5	0.2	0.002	76.467	< .001
		0.8	0.026	451.269	< .001
2,500	0.1	0.2	0.002	0.602	< .001
		0.8	0.026	0.592	< .001
	0.5	0.2	0.002	15.334	< .001
		0.8	0.026	12.936	< .001
5,000	0.1	0.2	0.002	0.442	< .001
		0.8	0.026	0.427	< .001
	0.5	0.2	0.002	11.108	< .001
		0.8	0.026	11.070	< .001

***True parameter value is 2.5**

Research Question Two

The second research question aims to measure the effect of the City Connect integrated student support model on student academic achievement in a single school district (District Z) using OLS, 2SLS-IV, Latent Instrumental Variable, and Gaussian copula regression models, with emphasis on comparing treatment effect estimates across methods. Additionally, by fitting both instrumental variable and instrument-free methods to real-world school lottery data, the second research question also seeks to illustrate the relationship between model parameters generated from instrument-free methods and a real-world, high-quality observed instrument from a RCT study. For the first portion of the second research question, the researcher analyzed both quasi-lottery and full lottery randomization data, comparing model parameter estimates both within and across lottery study designs. The second part of the second research question involved generating an optimal Latent Instrumental Variable instrument and comparing this to both the observed instrument from an RCT and the Gaussian copula treatment effect point estimate.

Lottery Study Impact of City Connects

Full lottery randomization study. The researcher first analyzed the data set containing only those students from the District Z lottery file who chose to submit a school preference list and thereby participate in the District Z school assignment process for kindergarten entry; the data set was then further refined to capture only those students who were assigned to a District Z City Connects school via lottery randomization (*data 2.b*). Students subjected to lottery randomization were identified by simulating the deferred acceptance algorithm $n = 100{,}000$ times and keeping only those students for which assignment to a City Connects school happened anywhere between approximately

500 to n – 500 times[11]. With these data, the researcher performed two cross-sectional analyses: 1) a kindergarten analysis (N=3,277); and 2) 3rd grade analysis (N=1,384). The kindergarten analysis was conducted in order to establish a baseline immediately post-randomization; the third grade analysis gives treatment effect estimates of the City Connects intervention for students receiving up to four years of the treatment. To generate causal effects of the treatment, OLS, 2SLS-IV, least squares Gaussian copula, and nonparametric Bayesian LIV regression models were fit to the reduced lottery data at both kindergarten and 3rd grade time points. The four model specifications are as follows:

Ordinary Least Squares:

$$Z\,RC/MCAS_{it} = \alpha_t + X_i'\beta + \beta_{CityConnects}CityConnects_{it} + \varepsilon_{it,} \quad (66)$$

Instrumental Variable Regression:

Second stage:

$$Z\,RC/MCAS_{it} = \alpha_{2t} + \textstyle\sum_j \delta_j d_{ij} + X_i'\beta + \beta_{CityConnects}\widehat{CityConnects}_{it} + \varepsilon_{it,}$$

(67A)

First stage:

$$CityConnects_{it} = \alpha_{1t} + \textstyle\sum_j \kappa_j d_{ij} + X_i'\Pi + \Pi_{Offer}Random\,Offer_i + \eta_{it,}$$

(67B)

Least Squares Gaussian copula Regression:

$$Z\,RC/MCAS_{it} = \alpha_t + X_i'\beta + \beta_{CityConnects}CityConnects_{it} + \beta_{CityConnects^*}CityConnects_{it}^* + \varepsilon_{it,} \quad (68)$$

[11] The frequency of assignment was smoothed such that it was rounded to the nearest hundredth, and thus the integers 500 and $n - 500$ were chosen because they would result in proportions to the nearest hundredths that would be between 0 and 1. The simple, and more general, idea is that for n simulations of the assignment algorithm, students are non-deterministically placed if assignment to a City Connects school happens anywhere between 1 to $n - 1$ times.

Bayesian Latent Instrumental Variable Regression:

$$Z\ RC/MCAS_{it} = \alpha_t + X_i'\beta + \beta_{CityConnects}CityConnects_{it} + \varepsilon_{it,} \quad (69A)$$

$$\alpha_t \sim N(\mu, \sigma^2),$$

$$\beta_i \sim N(\mu, \sigma^2) : \beta_i \in \beta,$$

$$\beta_{CCNX} \sim N(\mu, \sigma^2),$$

$$CityConnects_{it} = \theta_i + v_i, \quad (69B)$$

$$\theta_i \sim G,$$

$$G \sim DP(\alpha, G_0),$$

Where α_t are year effects and X is a design matrix containing a vector of 1's and the following student-level dummy covariates: gender (female=0, male=1); race (non-group membership=0, group membership=1); special education status (no = 0, yes = 1); free and reduced priced lunch status (no=0, yes=1); and bilingual status (no=0, yes=1). Furthermore, d_{ij} are the frequencies with which students were assigned to a City Connects school across the n simulation runs of the deferred acceptance algorithm, i.e., the DA propensity score; $Random\ Offer_i$ is a dummy indicator indicating random lottery offer; $CityConnects_{it}$ is the treatment dosage up until time point t; and $CityConnects_{it}^*$ is the inverse cumulative distribution of $CityConnects_{it}$.

We note that $CityConnects_{it}$ varies across cross-sectional analyses due to the amount of time from randomization to outcome reporting varying across analyses; so for the kindergarten analysis, $CityConnects_{it}$ represents the number of school months spent in the City Connects intervention up until report card reporting, as a yearly measure of dosage immediately following kindergarten randomization is severely restricted[12]. For the 3rd grade analysis, $CityConnects_{it}$ represents the number of years spent in the City Connects intervention up until test date. Furthermore, we note that there were two sets of outcome variables, one for each cross-sectional analysis. For the kindergarten analysis, the outcomes were math and reading report card scores standardized by subject, grade, and school year, denoted $Z\ RC_{it}$. Report card scores were used because standardized assessment data is unavailable for grades prior to 3rd grade. For the grade 3 analysis, the outcomes were scores on the math and English Language Arts (ELA) sections of a state-administered standardized assessment. The scores, denoted $Z\ MCAS_{it}$, have been standardized by subject, grade, and school year. Table 27 provides City Connects treatment effect estimates across grades and methods.

[12] To be further sure of the Kindergarten City Connects effect immediately post-randomization, the researcher also ran OLS and IV analyses using a dichotomous City Connects dose variable, examining if the substantive results from these analyses differed from kindergarten analyses using months spent in City Connects as the treatment variable. The substantive findings from this analysis were the same as those reported in Table 28, giving validity to the findings reported. This additional analysis can be found in the Appendix section.

Table 27.

Impact of City Connects intervention

		OLS		IV		Gaussian copula		Bayes LIV		
Grade	*Subject*	$\hat{\beta}(S.E.)$	*p*-value	$\hat{\beta}(S.E.)$	*p*-value	$\hat{\beta}(S.E.)$	*p*-value	$\hat{\beta}(S.E.)$	95% Credible Interval	λ
K	Math	<.01 (.01)	0.99	.01 (.02)	0.55	>-.01 (.01)	0.77	>-.01 (.01)	(-.02, .01)	0.53
	Reading	>-.01 (.01)	0.78	.01 (.02)	0.59	-.01 (.01)	0.50	-.01 (.01)	(-.02, .01)	0.53
3^{rd}	Math	.05 (.02)	< 0.01***	.04 (.06)	0.52	.06 (.03)	.04**	.04 (.02)	(.01, .08)**	0.67
	ELA	.02 (.02)	0.23	<.01 (.06)	0.96	.01 (.03)	0.41	.01 (.02)	(-.02, .05)	0.66

* denotes $p < 0.10$; ** denotes $p < 0.05$; *** denotes $p < .01$

Results from Table 27 show that all methods give surprisingly similar results, providing empirical validity evidence for the proposed instrument-free methods. For the kindergarten analysis, all methods indicate that there is no statistically significant difference in standardized report card scores between City Connects students and non-City Connects students. This was expected, as these outcomes come very shortly after the randomization process, and thus the treatment group has received minimal dosage at the time of outcomes reporting. Moving to 3rd grade, we see that there is no statistically significant difference in performance on the English Language Arts portion of the state test between City Connects students and non-City Connects students. For the mathematics portion of the state test, however, we see that three out of the four methods indicate a statistically significant difference in performance between City Connects and non-City Connects students, with City Connects students performing significantly higher than their non-City Connects peers. Furthermore, the estimates across all four methods are strikingly similar ($\hat{\beta}^{OLS} = 0.05$; $\hat{\beta}^{IV} = 0.04$; $\hat{\beta}^{GC} = 0.06$; $\hat{\beta}^{LIV} = 0.04$). We note that the IV and LIV estimates of the City Connects treatment effect are virtually identical. In interpreting the findings, regression estimates imply that every year of the City Connects intervention causes a $.04\sigma$ to $.06\sigma$ increase in mathematics measured on a state test relative to the counterfactual. Therefore, students randomized into the City Connects intervention at grade kindergarten and receiving the intervention through 3rd grade score $.16\sigma$ to $.24\sigma$ higher in mathematics on a state test than students not receiving the City Connects intervention during the same time period. In terms of practical significance, this is a small, positive effect for receiving the City Connects intervention between kindergarten and 3rd grade.

In examining the $\hat{\beta}(S.E.)$ column of Table 27, we see that the IV estimate is by far the least efficient, and thus the coefficient for this method is statistically non-significant. This result was also seen in previous simulation studies comparing IV regression with instrument-free methods, as findings suggested that IV is less efficient than instrument-free approaches when one makes use of anything but the true, perfect instrument, even if the observed instrument being used is valid and of high quality. Thus, such findings suggest that the instrument being used for this study is a high quality instrument but may not be not the true, perfect instrument[13]. Equally interesting is that the LIV λ estimate in Table 27, representing the probability of group membership, is what one would expect it to be given the lottery randomization taking place at kindergarten entry. For the school assignment process, students competing within lotteries have equal chances of being assigned to one of two groups (i.e., lottery offer versus lottery non-offer). In assessing λ, we see that the estimated probability of group membership is .53, reflecting nearly equal chances of being in the first of $m = 2$ groups for the simple Bayes LIV model. For grade 3 analyses, we see that the λ estimate changes, now reflecting unequal chances of group membership. However, we would expect this probability to change for later time points, as randomization takes place at kindergarten and students are not beholden to their lottery offers[14]; therefore students can, and often do, move around, especially as more time passes. Overall, the substantive

[13] The author expected this, as the instrument used for this study was a random lottery offer to a school that was coded as being a City Connects school *if it had ever been a City Connects school across a number of years*. Thus, the coding scheme is imperfect, albeit still fairly accurate and useful.
[14] This is essentially an issue of compliance, which could be addressed by an intention-to-treat analysis.

results and similarity of the estimates across methods from Table 27 provide strong empirical validity evidence for both the Gaussian copula and LIV approaches.

Endogeneity bias. The researcher examined evidence for endogeneity bias across methods at both kindergarten and 3rd grade time points. To investigate endogeneity bias under the instrumental variable approach, the researcher performed the Wu-Hausman test, which is essentially a test of the difference between the OLS estimate and IV estimate under the null hypothesis that both estimates are consistent and no endogeneity bias is present. Table 28 provides results for the Hausman test across grades.

Table 28.

IV Endogeneity test

Grade	Subject	Wu-Hausman statistic	*p*-value
K	Math	0.10	0.76
	Reading	0.04	0.85
Grade 3	Math	0.01	0.95
	Reading	0.01	0.93

At both kindergarten and grade 3 analyses, we note that endogeneity bias does not appear to be present, as we fail to reject the null hypothesis for all Wu-Hausman tests in Table 28. This finding is furthermore supported by the noted similarity between the OLS and IV regression coefficients reported in Table 27. However, it is important to note that the IV Wu-Hausman test for endogeneity assumes a valid instrument.

The researcher subsequently examined for evidence of endogeneity bias using instrument-free approaches. For the Gaussian copula regression, endogeneity bias is indicated by significant results from the Hausman test, which is simply a *t*-test on the

Gaussian copula Control Function term, $CityConnects^{*}$, in the least squares specification of the model given by Equation 68. The *t*-test is calculated as follows:

$$\frac{\hat{\beta}_{CityConnects^{*}}}{\sqrt{Var(\hat{\beta}_{CityConnects^{*}})}} \qquad (70)$$

For the Bayesian LIV approach, presence of endogeneity bias is determined from the 95% credible interval for the ρ parameter estimate, which captures the correlation between the endogenous regressor and the structural error term. Table 29 gives endogeneity test results across the instrument-free methods.

Table 29.

Instrument-free endogeneity tests

Model		**Gaussian copula**		**LIV**	
Grade	Subject	$\hat{\beta}_{CityConnects^{*}}(S.E.)$	p-value	$\hat{\rho}$ $(Posterior\ S.D.)$	95% CI
K	Math	.012 (.031)	0.694	.280 (.060)	(.165, .394)
	Reading	.014 (.029)	0.632	.273 (.058)	(.159, .392)
3	Math	-.018 (.047)	0.707	.049 (.072)	(-.098, .193)
	ELA	-.009 (.045)	0.837	.035 (.071)	(-.110, .167)

From Table 29, we see that findings from the Hausman test on the Gaussian copula Control Function agree with findings from the IV Wu-Hausman test, suggesting no endogeneity bias ($p > 0.05$). Interestingly, the LIV ρ estimate and its corresponding 95% credible interval for the Kindergarten analyses do not agree with previous findings and instead suggest endogeneity bias, with endogeneity ranging from .17 to .39. Using Cohen's conventions for small, medium, and large effects for Pearson's *r*, such values suggest small to moderate endogeneity bias. However, by the 3rd grade analyses, the results from all methods agree, with endogeneity tests across all three methods suggesting the presence of very little to no endogeneity bias.

Model diagnostics. Both instrument-free methods assume non-normality of the endogenous regressor, as model identifiability breaks down under violations of this assumption (Papies et al., 2017; Park & Gupta, 2012). Therefore, the researcher empirically assessed the endogenous regressor for non-normality by using histogram graphical displays of the treatment variable and performing the Shapiro-Wilk test for normality. At both kindergarten and 3rd grade time points, we note a highly non-normal distribution in the histogram display for treatment. Additionally, results from the Shapiro-Wilk test, where the null hypothesis is that the observed sample came from a normal distribution, indicate non-normality of the treatment variable.

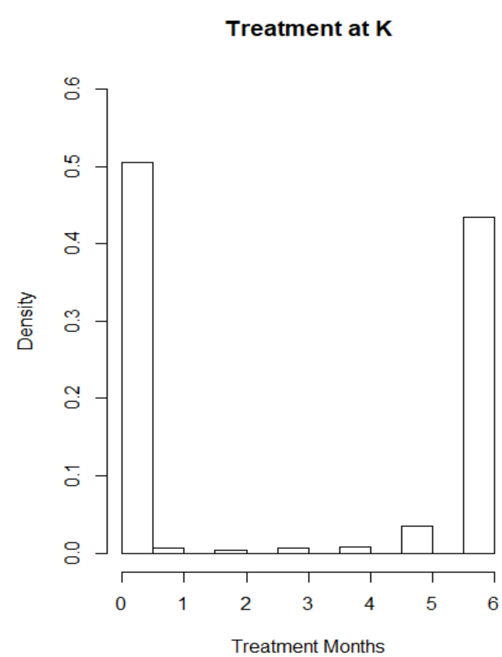

Figure 3. Histogram of Kindergarten treatment

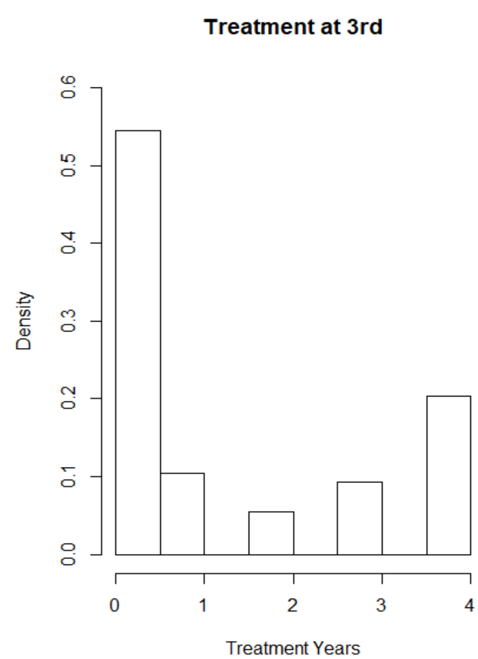

Figure 4. Histogram of 3rd grade treatment

Table 30.

Normality test for treatment variable

Shapiro-Wilk test of Normality		
Grade	W	*p*-value
K	0.66	< .01
3	0.72	< .01

In estimating the Bayesian LIV model, it is important to check for convergence to the target distribution, namely the posterior distribution (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012). The researcher examined for evidence of convergence of the Markov chains via trace plots of parameter estimates. Trace plots show the sampled parameter values taken by each chain for the duration of the chain (Lunn et al., 2012). In examining for convergence, the simple idea is that we start multiple chains (for these analyses there are three) and examine if they come together and begin to behave similarly (Lunn et al., 2012). We see from Figures 5 to 8 that all chains appear to converge to the posterior distribution after approximately 350 iterations (where burn-in is 300). In other words, we can draw a straight line through the three chains and see them similarly move around this line, i.e., parameter value. Thus, we note that all chains appear non-problematic across models and provide evidence for convergence[15].

[15] Trace plots are provided for City Connects treatment effects only. This was done to limit the amount of output. Trace plots for all parameters across all analyses can be found in the appendix section.

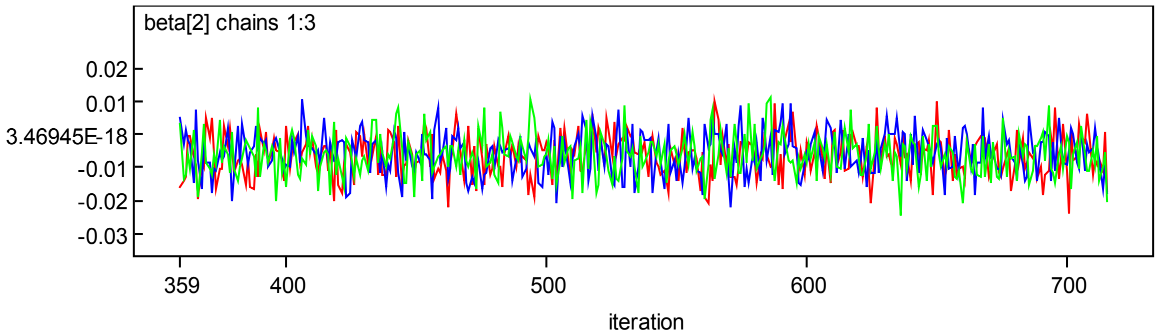

Figure 5. Trace Plot for Kindergarten City Connects ELA Effect

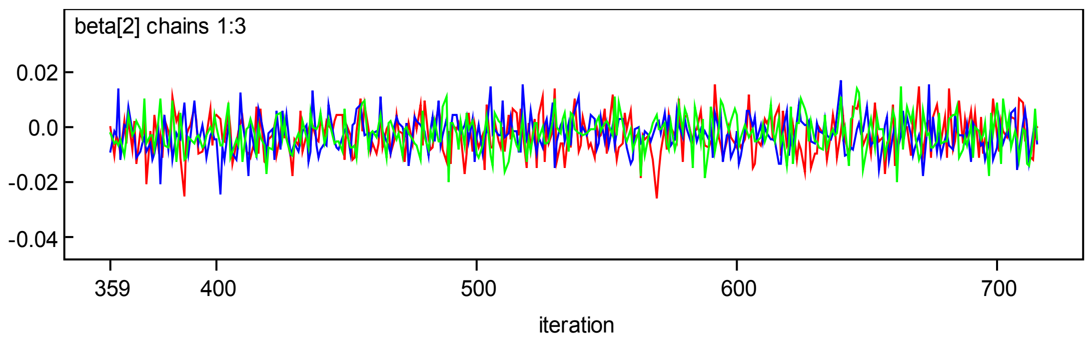

Figure 6. Trace Plot for Kindergarten City Connects Math Effect

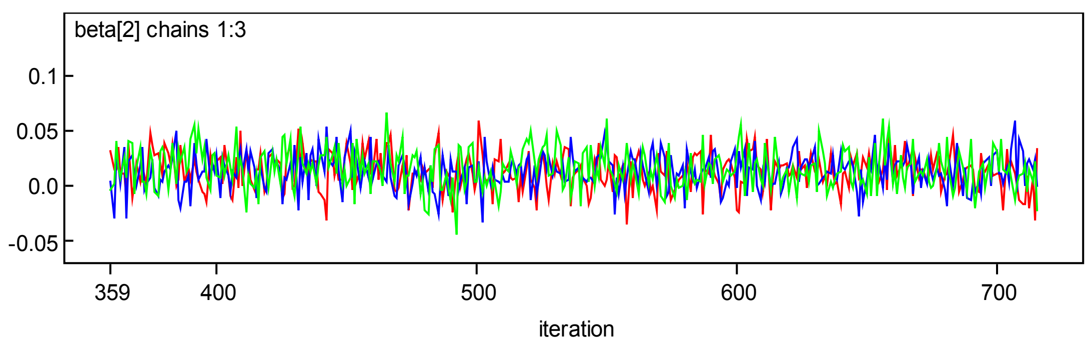

Figure 7. Trace Plot for Grade 3 City Connects ELA Effect

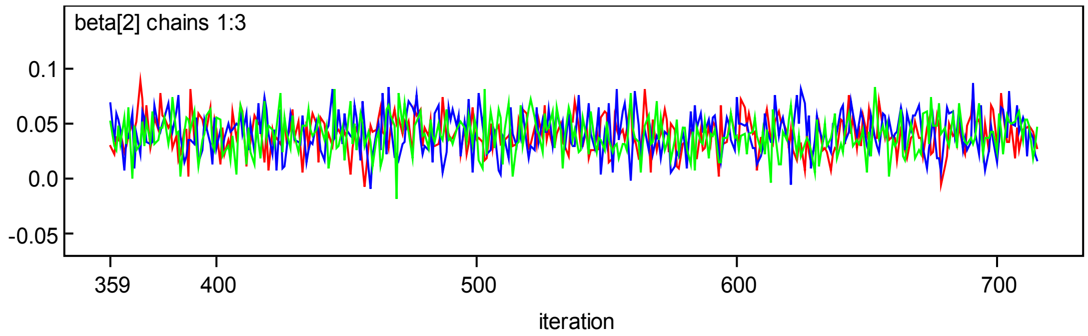

Figure 8. Trace Plot for Grade 3 City Connects Math Effect

Additionally, we note from Tables 31 and 32 that all MC error estimates are less than 5% of the corresponding posterior standard deviation estimate, providing evidence for sufficient iterations. Lastly, the researcher examined the Gelman-Rubin (1992) $\hat{R}$ statistic across models, which is the ratio of between-chain variability to within-chain variability. The $\hat{R}$ statistic is a MCMC convergence statistic for which the general rule of thumb is that values close to 1 indicate convergence and values above 1.1 indicate inadequate convergence. Note that for convergence, the $\hat{R}$ values for all parameters must be less than 1.1 (Brooks, Gelman, Jones, & Meng, 2011). From Tables 31 and 32, we see that all $\hat{R}$ statistics are very close to 1 in value[16].

[16] Given the similarity between math and reading/ELA samples, and to limit the amount of in-text output, MC error and $\hat{R}$ statistics are given for math analyses only; additional MC error and $\hat{R}$ statistics for reading/ELA analyses are provided in the appendix

Table 31.

MC Error and $\hat{R}$ statistics for kindergarten math

Parameter	**S.D.**	**MC Error**	$\hat{R}$
Intercept	0.077	0.003	1.04
City Connects	0.007	< .001	1.01
Year 07	0.069	0.002	1.03
Year 08	0.071	0.003	1.03
Year 09	0.070	0.002	1.01
Year 10	0.070	0.002	1.02
year 11	0.070	0.002	1.02
Year 12	0.067	0.002	1.01
Year 13	0.082	0.002	1.01
Male	0.035	0.001	1.00
Black	0.071	0.002	1.01
Hispanic	0.067	0.002	1.01
Asian	0.081	0.002	1.01
Mixed	0.131	0.004	1.00
Special Ed. 1	0.144	0.005	1.00
Special Ed. 2	0.088	0.003	1.00
Special Ed. 3	0.251	0.008	1.00
Free Lunch	0.052	0.002	1.00
Reduced Lunch	0.095	0.003	1.00
ELL	0.047	0.001	1.00
λ_1	0.010	< .001	1.00
λ_2	0.010	< .001	1.00
π_1	0.008	< .001	1.00
π_2	0.008	< .001	1.00
ρ	0.060	0.002	1.00
σ_1	0.023	< .001	1.00
σ_2	0.002	< .001	1.00

Table 32.

MC Error and $\hat{R}$ statistics for grade 3 math

Parameter	S.D.	MC Error	$\hat{R}$
Intercept	0.098	0.003	1.00
City Connects	0.018	< .001	1.00
Year 10	0.079	0.002	1.00
year 11	0.085	0.003	1.00
Year 12	0.080	0.002	1.00
Year 13	0.078	0.002	1.00
Male	0.049	0.001	1.00
Black	0.093	0.003	1.00
Hispanic	0.090	0.003	1.00
Asian	0.103	0.004	1.00
Mixed	0.190	0.006	1.00
Special Ed. 1	0.161	0.005	1.00
Special Ed. 2	0.114	0.004	1.00
Special Ed. 3	0.210	0.006	1.00
Free Lunch	0.075	0.003	1.00
Reduced Lunch	0.131	0.004	1.00
ELL	0.070	0.002	1.00
λ_1	0.014	< .001	1.00
λ_2	0.014	< .001	1.00
π_1	0.021	< .001	1.00
π_2	0.031	0.001	1.00
ρ	0.070	0.002	1.00
σ_1	0.033	0.001	1.00
σ_2	0.012	< .001	1.00

Quasi-lottery study. The researcher analyzed the data set containing all District Z students from the lottery file for which demographic and student outcomes data are also available, regardless of whether a student chose to opt out of the District Z school assignment process or participate (*data 2.a*). Given that the quasi-lottery data contains students that did not participate in the school lottery assignment process, an instrumental variable analysis using a random lottery offer as an instrument cannot be used for this sample; however, the benefit of instrument-free approaches is that they do not rely on an observed instrument from a lottery mechanism, and thus these approaches can be used for addressing endogeneity bias in the extended sample. The researcher fit OLS, Gaussian copula, and Bayesian Latent Instrumental Variable regression models to the extended quasi-lottery sample at kindergarten and 3rd grade time points, as was done for the full lottery randomization sample. Table 33 provides results from across the three methods.

Examining Table 33, we once again see that all methods give surprisingly similar results. For the kindergarten analysis, all three methods indicate that City Connects students have slightly lower standardized report card scores than their non-City Connects peers. Furthermore, all methods suggest that this difference is statistically significant for both Reading and mathematics. We note, however, that such findings are for the kindergarten grade level and thus reflect minimal dosage; in other words, we can view results from this analysis as providing a starting point or baseline comparison.

Moving to 3rd grade, we see that there is no statistically significant difference in performance on the English Language Arts portion of the state test between City Connects students and non-City Connects students. For the mathematics portion of the state test, we see that instrument-free methods indicate a statistically significant difference in performance between City Connects and non-City Connects students, with City Connects students performing significantly higher than their non-City Connects peers. Furthermore, the estimates across all three methods are very similar ($\hat{\beta}^{OLS} = 0.03$; $\hat{\beta}^{GC} = 0.04$; $\hat{\beta}^{LIV} = 0.04$). We note that the Gaussian copula and LIV estimates of the City Connects treatment effect are exactly the same. Moreover, these results are similar to the estimates for the randomization sample, suggesting that the effect of City Connects is similar across samples and that there may not be considerable lottery selection bias. In interpreting the findings, regression estimates imply that every year of the City Connects intervention causes a $.03\sigma$ to $.04\sigma$ increase in mathematics on a state test relative to the counterfactual for the general sample. Therefore, students entering into the City Connects intervention at grade kindergarten and receiving the intervention through 3rd grade score $.12\sigma$ to $.16\sigma$ higher in mathematics on a state test than students not receiving the City Connects intervention during the same time period. In terms of practical significance, this is once again a small, positive effect for receiving the City Connects intervention between kindergarten and 3rd grade.

In examining the LIV λ estimate in Table 33, representing the probability of group membership, we once again see what one would expect given the sample under consideration. For the analysis of the randomization sample in the previous section, the LIV λ estimates were roughly .50, reflecting the randomization process taking place. Contrastingly, for this kindergarten analysis, we now see that kindergarten LIV λ estimates are far from .50, and no longer reflect nearly equal chances of group membership. However, this change makes sense, as we are now analyzing the full sample of students, which includes both lottery and non-lottery participants, and we therefore no longer have a randomized control trial (RCT).

Table 33.

Full sample analyses

		OLS		Gaussian copula		Bayes LIV		
Grade	*Subject*	$\hat{\beta}(S.E.)$	*p*-value	$\hat{\beta}(S.E.)$	*p*-value	*Subject*	95% Credible Interval	λ
K	Math	-.01 (< .01)	< .01***	-.01 (< .01)	< .01***	-.02 (< .01)	(-.02, -.01)**	0.71
	Reading	-.01 (< .01)	< .01***	-.01 (< .01)	< .01***	-.02 (< .01)	(-.02, -.01)**	0.70
3	Math	.03 (.01)	< .01***	.04 (.01)	< .01***	.04 (.01)	(.03, .05)**	0.86
	ELA	> -.01 (.01)	0.36	< .01 (.01)	0.88	-.01 (< .01)	(-.02, < .01)	0.89

* denotes $p < 0.10$; ** denotes $p < 0.05$; *** denotes $p < .01$

Note: Standard errors are clustered by school for OLS and Gaussian copula methods

As noted previously, estimates of the City Connects intervention effect were similar across lottery samples. Specifically, the estimates of the City Connects effect for the quasi-lottery study are only slightly smaller than those for the randomization sample. Such similarity suggests that lottery participation does not have a huge impact on the effectiveness of the City Connects intervention, i.e., lottery selection bias may not be of great concern. However, the similarity of the regression coefficients across lottery designs is far from definitive proof that students and their families who self-select into the kindergarten lottery process do not differ systematically from those students and families who choose to opt out. Therefore, the author seeks to further investigate the differences between the two samples. To do so, the author examined descriptive statistics for the following available demographic variables across the two samples: gender; race; reduced and free priced lunch; ELL status; and immigration status. In addition to examining sample percentages, the author calculated standardized mean differences to make claims about important covariate imbalance between lottery and non-lottery samples. A cutoff of .20 was used for determining imbalance. Table 34 provides descriptive statistics for lottery and non-lottery samples.

We see from Table 34 that the lottery sample, which contains only those students who participated in the lottery assignment process, is similar to the non-lottery sample in numerous ways. Specifically, the lottery sample is similar to the non-lottery sample in regards to gender, special education status, ELL status, immigration status, and reduced lunch status, which serves as one proxy for socioeconomic status. However, we do notice a few differences between the two groups. Although below the .20 cutoff, students receiving free lunch are noticeably more prevalent in the non-lottery sample, whereas

Hispanic students are more prevalent in the lottery sample. We notice the greatest imbalance for African-American students, with the corresponding SMD statistic exceeding the .20 cutoff. From this, we can say that African-American students and their families seem less likely to participate in the lottery process. Such findings may limit the generalizability of IV findings based on lottery participation to larger student populations. However, given that instrument-free methods do not rely on an observed instrument, the researcher was able to estimate exogenous City Connects treatment effects for samples including both groups, and thus one is able to make broader claims about the causal effect of the intervention.

Table 34.

Sample demographics at kindergarten entry point

Variable	Non-lottery Sample	Lottery Sample	SMD
Male	52%	51%	0.01
Black	37%	26%	0.24
Asian	7%	9%	0.07
Hispanic	41%	49%	0.17
Mixed	2%	2%	0.03
Sped	7%	7%	< .01
Reduced lunch	4%	4%	0.01
Free lunch	83%	76%	0.18
ELL	20%	23%	0.07
Foreign Born	9%	7%	0.08

Endogeneity bias. Using the full sample, the researcher once again examined for evidence of endogeneity bias across instrument-free methods at both kindergarten and 3rd grade time points. Table 35 gives endogeneity test results across the instrument-free methods.

From Table 35, we see that the LIV ρ estimate and its corresponding 95% credible interval for the Kindergarten analyses using the full sample suggests endogeneity bias, with endogeneity ranging from .19 to .24. Using Cohen's conventions for small, medium, and large effects for Pearson's *r*, such values suggest small endogeneity bias. Interestingly, findings from the Gaussian copula Hausman test once again do not agree with the LIV estimates. For the Gaussian copula regression, results from the Hausman test on the copula Control Function term suggest no endogeneity bias ($p > 0.05$). However, by the time we reach 3rd grade analyses, the results from all methods agree, with endogeneity tests across the two methods suggesting no presence of endogeneity bias.

Table 35.

Instrument-free endogeneity tests

Model		**Gaussian copula**		**LIV**	
Grade	Subject	$\hat{\beta}_{CityConnects^*}(S.E.)$	p-value	$\hat{\rho}$ $(Posterior\ S.D.)$	95% CI
K	Math	>-.01 (.01)	0.66	.22 (.01)	(.19, .24)
	Reading	-.01 (.01)	0.27	.21 (.01)	(.19, .23)
3	Math	-.01 (.01)	0.35	-.02 (.01)	(-.05, .01)
	ELA	-.01 (.01)	0.11	.01 (.01)	(-.02, .03)

Model diagnostics. The researcher empirically assessed the endogenous regressor for non-normality by using histogram graphical displays of the treatment variable. At both kindergarten and 3rd grade time points, we note a highly non-normal distribution in the histogram display for treatment. Given the much larger sample size, and the tendency of the Shapiro-Wilk test to over-reject the null hypothesis in large samples, a Shapiro-

Wilk test for normality was not performed with the full sample, and instead the author relied solely on graphical displays of the treatment variable.

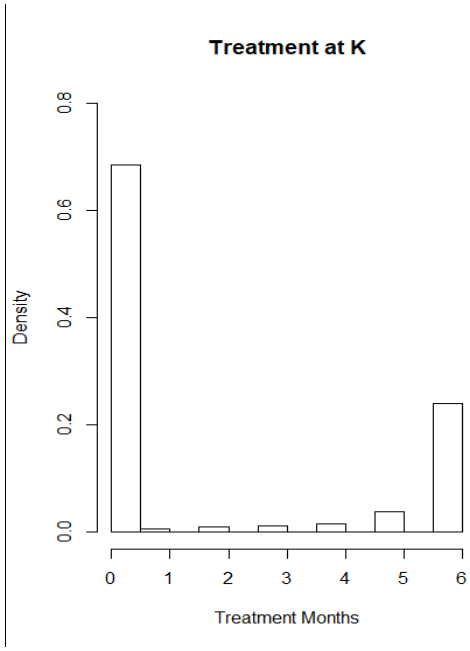

Figure 9. Histogram for treatment at kindergarten

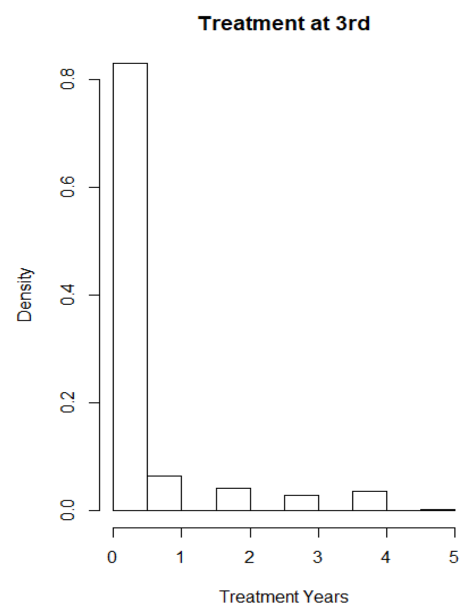

Figure 10. Histogram for treatment at 3rd grade.

Convergence of the Markov chains was assessed via trace plots of parameter estimates. Given the much larger sample sizes, and to help ensure convergence, the number of iterations for the full sample analyses was set to be substantially larger than they were for previous analyses, with the value now being $n_{iterations}$ = 10,000. We see from Figures 11 to 14 that all chains appear to converge to the posterior distribution.

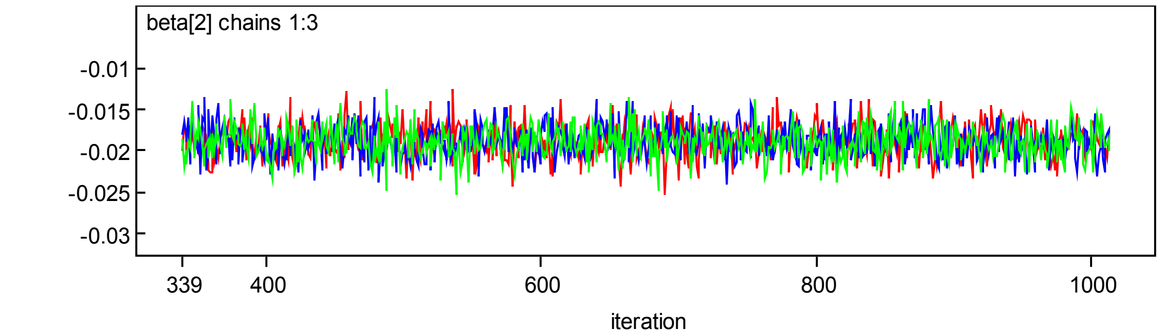

Figure 11. Trace Plot for Full sample Kindergarten City Connects ELA Effect

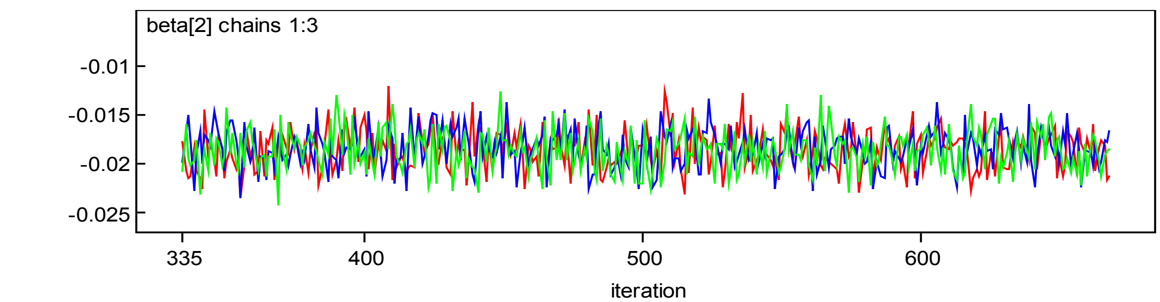

Figure 12. Trace Plot for Full Sample Kindergarten City Connects Math Effect

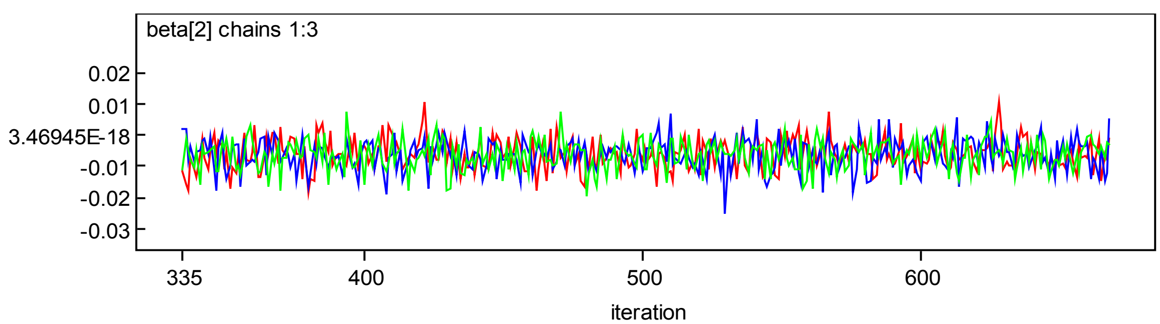

Figure 13. Trace Plot for Full Sample Grade 3 City Connects ELA Effect

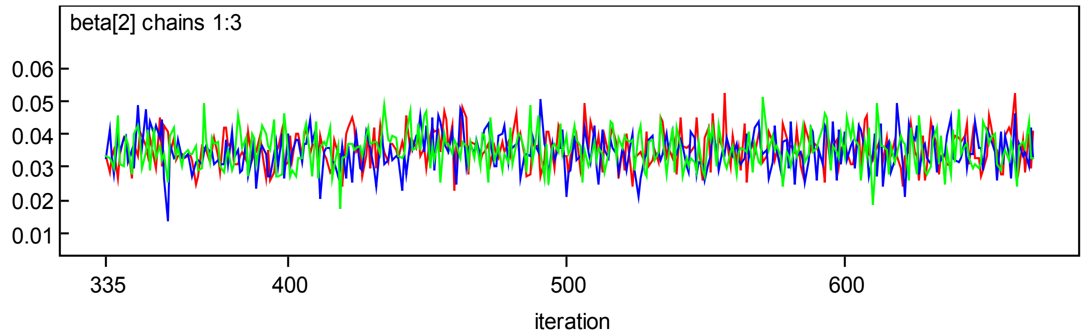

Figure 14. Trace Plot for Full Sample City Connects Math Effect

Additionally, we note from Tables 36 and 37 that all MC error estimates are once again less than 5% of the corresponding posterior standard deviation estimate, providing evidence for sufficient iterations. Additionally, from Tables 36 and 37, we note that the Gelman-Rubin $\hat{R}$ statistics are all very close to 1 in value.

Table 36.

MC Error and $\hat{R}$ statistics for kindergarten math full sample

Parameter	**S.D.**	**MC Error**	$\hat{R}$
Intercept	0.022	< .001	1.00
City Connects	0.002	< .001	1.00
Year 05	0.023	< .001	1.00
Year 06	0.022	< .001	1.00
Year 07	0.023	< .001	1.00
Year 08	0.023	< .001	1.00
Year 09	0.022	< .001	1.00
Year 10	0.022	< .001	1.00
year 11	0.022	< .001	1.00
Year 12	0.023	< .001	1.00
Year 13	0.025	< .001	1.00
Male	0.010	< .001	1.00
Black	0.017	< .001	1.00
Hispanic	0.017	< .001	1.00
Asian	0.023	< .001	1.00
Mixed	0.036	< .001	1.00
Special Ed. 1	0.045	0.001	1.00
Special Ed. 2	0.022	< .001	1.00
Special Ed. 3	0.058	0.002	1.00
Free Lunch	0.015	< .001	1.00
Reduced Lunch	0.030	< .001	1.00
ELL	0.013	< .001	1.00
λ_1	0.003	< .001	1.00
λ_2	0.003	< .001	1.00
π_1	0.003	< .001	1.00
π_2	0.004	< .001	1.00
ρ	0.014	< .001	1.00
σ_1	0.007	< .001	1.00
σ_2	0.001	< .001	1.00

Table 37.

MC Error and $\hat{R}$ statistics for grade 3 math full sample

Parameter	S.D.	MC Error	$\hat{R}$
Intercept	0.025	0.001	1.01
City Connects	0.005	< .001	1.00
Year 06	0.023	< .001	1.00
Year 07	0.023	< .001	1.00
Year 08	0.024	< .001	1.00
Year 09	0.024	< .001	1.00
Year 10	0.023	< .001	1.00
year 11	0.024	< .001	1.00
Year 12	0.023	< .001	1.00
Year 13	0.023	< .001	1.00
Male	0.011	< .001	1.00
Black	0.019	< .001	1.00
Hispanic	0.018	< .001	1.00
Asian	0.024	< .001	1.00
Mixed	0.043	0.001	1.00
Special Ed. 1	0.042	0.001	1.00
Special Ed. 2	0.022	< .001	1.00
Special Ed. 3	0.033	0.001	1.00
Free Lunch	0.019	< .001	1.00
Reduced Lunch	0.035	0.001	1.00
ELL	0.013	< .001	1.00
λ_1	0.002	< .001	1.00
λ_2	0.002	< .001	1.00
π_1	0.003	< .001	1.01
π_2	0.007	< .001	1.00
ρ	0.014	< .001	1.00
σ_1	0.007	< .001	1.00
σ_2	0.002	< .001	1.00

Further examination of the 3rd Grade City Connects Math Effect. In light of the previous findings, the author wished to further investigate the impact of City Connects on 3rd grade math achievement by fitting models that account for the school-to-school variability and explicitly model the multilevel structure of the data. Given the clustering of students within schools, and the likely presence of treatment effect heterogeneity, modeling approaches that incorporate school-level information arguably provide more valid inferences about the intervention impact. To explore this, fixed effect OLS and Gaussian copula regression and a random intercept nonparametric Bayesian LIV regression were fit to the full district sample including nonrandomized students. Modeling specifications from Equations 66 to 69 were extended as follows:

Ordinary Least Squares:

$$Z\,MCAS_{ijt} = \gamma_j + \alpha_t + X_i'\beta + \beta_{CityConnects}CityConnects_{ijt} + \varepsilon_{ijt,} \quad (71)$$

Least Squares Gaussian copula Regression:

$$Z\,MCAS_{ijt} = \gamma_j + \alpha_t + X_i'\beta + \beta_{CityConnects}CityConnects_{ijt} + \beta_{CityConnects^*}CityConnects^*_{ijt} + \varepsilon_{ijt,} \quad (72)$$

Random Intercept Bayesian Latent Instrumental Variable Regression:

$$Z\,MCAS_{ijt} = \gamma_j + \alpha_t + X_i'\beta + \beta_{CityConnects}CityConnects_{ijt} + \varepsilon_{ijt,} \quad (73A)$$

$$\gamma_j \sim N(\beta_0, \tau^2)$$

$$\beta_0 \sim N(\mu, \sigma^2)$$

$$\alpha_t \sim N(\mu, \sigma^2),$$

$$\beta_i \sim N(\mu, \sigma^2) : \beta_i \in \beta,$$

$$\beta_{CityConnects} \sim N(\mu, \sigma^2),$$

$$CityConnects_{ijt} = \Theta_i + \upsilon_i, \quad (73B)$$

$$\Theta_i \sim G,$$

$$G \sim DP(\alpha, G_0),$$

Where the models are specified as they were before in Equations 66 to 69 but now include γ_j, which is the fixed school effect in the OLS and Gaussian copula model and the random school intercept in the nonparametric Bayesian LIV model. Given that there are no predictors of the random intercept in the multilevel Bayesian LIV model, the estimates across models should be very similar. Table 38 provides results for the OLS, Gaussian copula, and random intercept Bayesian LIV model.

Table 38.

Fixed effect and random intercept models

		OLS		**Gaussian copula**		**Random Intercept Bayes LIV**		
Grade	*Subject*	$\hat{\beta}(S.E.)$	*p*-value	$\hat{\beta}(S.E.)$	*p*-value	*Subject*	95% CI	λ
3	Math	.07 (.01)	< .01***	.08 (.01)	< .01***	.07 (.01)	(.06, .09)**	0.86

Examining Table 38, we again see that all methods give very similar results. Given that there is no predictor for the level-two random intercept in the multilevel Bayesian LIV model, this is to be expected. Consistent with previous analyses, all three methods indicate a statistically significant difference in mathematics performance between City Connects and non-City Connects students, with City Connects students performing significantly higher than their non-City Connects peers. Furthermore, the estimates across all three methods are very similar ($\hat{\beta}^{OLS} = 0.07$; $\hat{\beta}^{GC} = 0.08$; $\hat{\beta}^{LIV} = 0.07$). We note that the OLS and LIV estimates of the City Connects treatment effect are identical. However, estimates from this set of analyses indicate a larger achievement gain than the estimates from previous analyses with simpler models. Such a difference indicates that clustering and between school variability matters, and there is likely treatment effect heterogeneity. In interpreting the OLS and Gaussian copula findings, regression estimates imply that every year of the City Connects intervention causes a .07σ to .08σ increase in mathematics on a state test relative to the counterfactual, controlling for between school variability. For the multilevel LIV model, results indicate that every year of the City Connects intervention causes a .07σ increase in mathematics on a state test relative to the counterfactual, controlling for the random effect of school. Therefore, students entering into the City Connects intervention at grade kindergarten and receiving the intervention through 3rd grade score .28σ to .32σ higher in mathematics on a state test than students not receiving the City Connects intervention during the same time period. In terms of practical significance, this is a notable positive effect for receiving the City Connects intervention between kindergarten and 3rd grade.

ICC. Before estimating the multilevel Bayesian LIV model, the researcher examined the intraclass correlation coefficient (ICC) to examine the proportion of variance in 3rd grade state test mathematics achievement that exists between schools (O'Dywer & Parker, 2014). By estimating the ICC, the researcher was able to assess the degree of statistical dependency in the data and determine the need for a multilevel modeling approach (O'Dwyer & Parker, 2014). Although there is no general rule for how large the ICC needs to be before multilevel modeling is justified, the researcher examined for an ICC that was not close to zero in value. The ICC for these data was found to be .14. In other words, 14% of the variance in 3rd grade state test mathematics achievement was due to between school variability.

Endogeneity evidence. The researcher once again examined for evidence of endogeneity bias across instrument-free methods. Table 39 gives endogeneity test results across the instrument-free methods.

Table 39.

Endogeneity bias for fixed effect and multilevel Gaussian copula and LIV model

Model		**Gaussian copula**		**Random Intercept Bayes LIV**	
Grade	Subject	$\hat{\beta}_{CityConnects^*}(S.E.)$	p-value	$\hat{\rho}\ (Posterior\ S.D.)$	95% CI
3	Math	-.01 (.01)	0.04	-.05 (.01)	(-.07, -.02)

Results from the Gaussian copula and random intercept Bayesian LIV model agree, and we see from Table 39 that both methods suggest the presence of minor endogeneity bias. Although both methods suggest that the correlation between the endogenous regressor and error term is statistically significant, the values for this correlation are very small and close to zero. Such a small degree of endogeneity bias will not bias OLS estimates very much, and this is reflected in the findings reported for Table 38, as the regression coefficients from OLS and the instrument-free methods were very similar.

LIV Model convergence. Convergence of the Markov chains was assessed via trace plots of parameter estimates. Given the complexity of the model, the number of iterations for the random intercept LIV model was set to be substantially larger than they were for previous analyses, with the value now being $n_{iterations} = 25{,}000$. We see from Figure 15 that the treatment effect chain appeared to converge to the posterior distribution. All other chains appeared to converge as well and appear in the appendix. Additionally, we note from Table 40 that all MC error estimates are once again less than 5% of the corresponding posterior standard deviation estimate, providing evidence for sufficient iterations. Additionally, from Table 40, we note that the Gelman-Rubin $\hat{R}$ statistics are all very close to 1 in value.

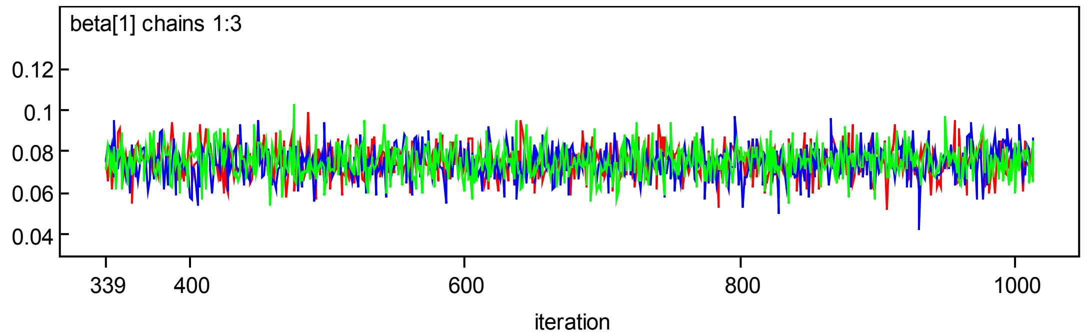

Figure 15. Trace Plot for Random Intercept LIV Model City Connects Math Effect

Table 40.

MC Error and $\hat{R}$ statistics for grade 3 fixed effects and multilevel models

Parameter	S.D.	MC Error	$\hat{R}$
Intercept	0.043	< .001	1.00
City Connects	0.008	< .001	1.00
Year 06	0.022	< .001	1.00
Year 07	0.022	< .001	1.00
Year 08	0.022	< .001	1.00
Year 09	0.022	< .001	1.00
Year 10	0.022	< .001	1.00
year 11	0.022	< .001	1.00
Year 12	0.022	< .001	1.00
Year 13	0.022	< .001	1.00
Male	0.010	< .001	1.00
Black	0.019	< .001	1.00
Hispanic	0.018	< .001	1.00
Asian	0.026	< .001	1.00
Mixed	0.043	< .001	1.00
Special Ed. 1	0.041	< .001	1.00
Special Ed. 2	0.022	< .001	1.01
Special Ed. 3	0.032	< .001	1.00
Free Lunch	0.019	< .001	1.00
Reduced Lunch	0.034	< .001	1.00
ELL	0.014	< .001	1.00
λ_1	0.002	< .001	1.00
λ_2	0.002	< .001	1.00
π_1	0.003	< .001	1.01
π_2	0.006	< .001	1.00
ρ	0.014	< .001	1.00
σ_1	0.006	< .001	1.00
σ_2	0.001	< .001	1.00

A comparison of instrument-free parameters to the observed instrument

Given that District Z school lottery admissions generate a stratified RCT, one very arguably is provided with a relevant and valid instrument when relying on a lottery offer variable. The author took advantage of the observed, high quality instrument available from the lottery RCT to further demonstrate the validity of the proposed instrument-free approaches. To do so, the researcher first produced a posterior distribution for the optimal Bayes LIV instrument by sampling from the full conditional distribution:

$$p(\tilde{z}_i \mid \tilde{z}_{-i}\, \beta, \Sigma, \alpha, G_0, b)\ , \qquad (74)$$

where b is the $n\ x$ 1 vector containing elements b_i, and where $b_i = (y_i, x_i)$. As was done for simulation study four, the optimal Bayes LIV instrument, $\tilde{z}$, was then calculated as the mean of this posterior distribution, rounded to the nearest integer. The rounded posterior mean produces an observed instrument from the estimated Bayes LIV model, allowing for comparison of the estimated latent instrument produced by the LIV model to the observed lottery instrument. The author then examined the degree to which the LIV instrument correlated with the lottery instrument and how well the estimated LIV instrument recreated the lottery offer via classification accuracy. Additionally, the researcher examined the degree to which the estimated LIV instrument correlated with the endogenous regressor (i.e., relevance). All comparative analyses were conducted at the kindergarten time point, as that is when the randomly generated lottery offer is produced. Table 41 provides correlations and cross-classification accuracies.

Table 41.

LIV instrument correlation analyses

	Optimal Bayes LIV Instrument Correlations		
Subject	$\bar{\rho}_{\tilde{z},lottery\ offer}$	**% correctly classified**	$\bar{\rho}_{\tilde{z},x}$
Math	.74	86%	.99
Reading	.73	86%	.99

Table 41 provides strong empirical validity evidence for the LIV approach, as we see that the estimated LIV instrument is strongly correlated with the observed, high quality instrument ($\bar{\rho}_{\tilde{z},lottery\ offer}$ = .74, = .73). Given the previously noted imperfect coding scheme used for the lottery instrument, we note that the observed instrument is not perfect either, and thus the correlation between the estimated LIV instrument and lottery instrument could be slightly attenuated due to this fact. However, we note that the observed lottery instrument is arguably still of very high quality, and such a strong correlation between the LIV instrument and observed lottery instrument speaks to the power of the instrument-free LIV method. Additionally, we note that the estimated LIV instrument correctly matches the lottery offer for 86% of the observations. Moreover, the estimated LIV instrument is highly relevant, producing a nearly perfect correlation with the endogenous regressor at $\bar{\rho}_{\tilde{z},x}$ = .99.

The researcher subsequently performed a 2SLS-IV regression using the estimated LIV instrument (denoted 2SLS-LIV) and compared this with 2SLS-IV using the observed lottery instrument and least squares Gaussian copula regression. Table 42 provides comparative results.

Table 42.

2SLS-LIV comparative results

		IV		**Gaussian copula**		**2SLS-LIV**	
Grade	*Subject*	$\hat{\beta}(S.E.)$	p-value	$\hat{\beta}(S.E.)$	p-value	$\hat{\beta}(S.E.)$	p-value
K	Math	.011 (.018)	0.550	-.003 (.011)	0.765	-.003 (.006)	0.633
	Reading	.009 (.017)	0.593	-.007 (.010)	0.503	-.006 (.006)	0.314

We note from Table 42 that substantive findings across all three methods are the same, and the City Connects intervention has no statistically significant impact immediately post-randomization. Furthermore, we see that the 2SLS-LIV and Gaussian copula estimates are nearly identical. Estimates from the IV regression differ from the instrument-free method estimates but are also close to zero. Given the proximity to zero of all estimates and the non-statistical significance, all differences between estimates are deemed trivial. In sum, we see that the LIV model can be estimated in two ways, and the method produces an estimated latent instrument that is strongly correlated with an observed instrument from an RCT, thereby producing findings consistent with those from other empirically sound approaches.

CHAPTER 5: DISCUSSION

Summary of Findings

City Connects is an integrated student support model offering student support in high-poverty, urban schools. Given that City Connects does not randomly assign students to receive the intervention, the consideration of statistical methods for dealing with endogeneity bias is important. This dissertation research explored the utility of instrument-free methods for addressing endogeneity bias. Specifically, the author investigated two research questions:

1) How does estimation performance under the two-stage least squares IV (2SLS-IV) approach, the Latent Instrumental Variable (LIV) approach, and the least squares Gaussian copula approach compare across a range of research conditions involving endogeneity bias?;
2) Using data from a real-world school lottery study examining the effect of the City Connects model of integrated student support, how do treatment effect estimates compare under the traditional 2SLS-IV approach with simulation-based propensity scores, the Latent Instrumental Variable (LIV) approach, and the Gaussian copula approach? And, how do the model parameters generated by instrument-free approaches compare to the observed instrument?

The first research question was investigated via extensive simulation study, whereas the second research question involved the application of instrument-free methods to a real-world large-scale RCT.

Simulation Findings

Study 1. The simulation study research comprised six discrete studies. For simulation study one, the researcher investigated instrument-free method performance under the condition of exogeneity. Results demonstrated that both LIV and Gaussian copula methods are unbiased and produce estimates that closely resemble OLS estimates; however, OLS is far more efficient than the instrument-free approaches and therefore provides the best linear unbiased estimate. Such findings are in perfect accordance with the extant literature, as Gauss and Markov proved OLS to be the best linear unbiased estimate under the conditions of exogeneity and error having mean zero, and the functional form being correctly specified (Ebbes, 2004; Hueter, 2016).

Study 2. Simulation study two investigated the performance of instrumental variable and instrument-free methods under endogeneity arising from a linear regression specification. Park and Gupta (2012) previously investigated the performance of instrumental variable and Gaussian copula regression approaches under such a condition; however, the authors did not investigate the performance of the LIV regression approach under this condition. Therefore, by also considering the LIV regression approach, this research provides new comparative findings.

Findings from this study suggest that IV and Gaussian copula regression approaches correctly adjust for the endogeneity bias when endogeneity is specified to be purely correlational, i.e., the endogeneity is entirely captured by $\rho_{x,\varepsilon}$, and there is no exogenous variation in the endogenous regressor. These findings match those from Park and Gupta (2012), where the authors also found that the IV and Gaussian copula approaches successfully recovered the true parameter values. In addition, findings from this dissertation study demonstrated that the LIV approach produced inaccurate estimates under this condition, especially as the degree of endogeneity increased. Such findings suggest that additive separability in the endogenous regressor, i.e., the endogenous regressor can be split into two additive pieces, an endogenous and exogenous component, is a strong requirement of the LIV modeling approach.

Study 3. The researcher subsequently investigated the performance of instrumental variable and instrument-free methods under endogeneity arising from a LIV regression model, where there is now additive separability in the endogenous regressor and thus there exists exogenous variation. In other words, the endogenous regressor is now represented as $x = \theta + \nu$, where θ is the exogenous latent instrument and ν is an additive error term that is endogenous. Park and Gupta (2012) previously compared the Gaussian copula regression approach with the LIV regression approach under this condition; however, their simulation research was based only on a simple linear regression through the origin (RTO) LIV model, where the intercept was omitted and there was a single slope. Therefore, this research extended their simulation work by generating endogenous data from three LIV models: a RTO LIV model; a full LIV model with intercept and slope; and a full LIV model with intercept and multiple slopes.

Simulation results for the RTO LIV model matched those from Park and Gupta (2012), as all three IV, Gaussian copula, and LIV modeling approaches produced unbiased, highly accurate estimates. However, once the author considered fuller parameterizations and simulated data from LIV models with both intercept and slope(s), the simulation results changed. For both full LIV model specifications, the IV and LIV regression approaches still produced highly accurate, unbiased estimates of the true parameter value, as the author expected. Contrastingly, however, the Gaussian copula approach now produced inaccurate, biased results. Such findings produce new evidence suggesting that the Gaussian copula approach has difficulty correcting for endogeneity bias when the exogeneity requirement for the instrument holds and the dependence structure differs from what the modeling approach assumes.

Study 4. The simulation study four further investigated the performance of the LIV regression approach with endogenous data based on both the linear regression and LIV specification. The dissertation research results from simulation study two showed that LIV estimates became increasingly inaccurate when endogeneity was based on a linear regression specification, yet this inaccuracy was not statistically significant according to t_{bias} test statistics. To further determine if the LIV model produced biased estimates under endogeneity arising from a linear regression specification, the author generated an observed LIV instrument by estimating a nonparametric Bayesian LIV model. A 2SLS regression using the generated LIV instrument was then fit to simulated data from both a linear regression specification and a LIV model specification.

For endogenous data arising from a LIV model, the 2SLS regression using the LIV instrument produced estimates that were unbiased and highly accurate. Furthermore, the generated LIV instrument correlated nearly perfectly with the true instrument and had roughly zero correlation with the true error term. However, for endogenous data based on the linear regression specification, the 2SLS regression based on the generated LIV estimate produced inaccurate, statistically significantly biased results. The bias was now detectable due to the increased efficiency of the 2SLS regression approach using the generated LIV instrument. Moreover, the generated LIV instrument was only weakly correlated with the true instrument and now correlated with the true error term. Interestingly, the correlation between the LIV instrument and true error was roughly half that of the endogeneity specified for the data generating process. Thus, this evidence revealed that the LIV approach corrects for roughly only half of the endogeneity bias when the dependence structure is different from what is assumed for the modeling approach. Additionally, these findings confirm that additive separability is a strong requirement for the LIV modeling approach. Furthermore, when this finding is taken in conjunction with findings from simulation study three, results suggest that there is an important additional, and previously undiscussed, assumption of the two proposed instrument-free approaches. Specifically, the added assumption of the instrument-free methods is that the endogeneity can be represented by a certain dependence structure; moreover, we see that the dependence structure assumed for each instrument-free method strongly matters. These findings are new and highly informative for future research, which will be further discussed in sections to follow.

Study 5. Misspecification of the error term was investigated for simulation study five. For this study, the author considered uniform, *F*-distribution(8,5), and Chi-square(4) distributions for the true structural error term. This research extends the work of Ebbes (2009) and Park and Gupta (2012) by considering new distributions for the structural error term, as Park and Gupta only investigated a Uniform distribution for the error term under the Gaussian copula approach, and Ebbes considered Gamma, mixture, Chi-square(1) and *t*(3) distributions for the LIV model error (2004; 2009; 2012).

When the error term was misspecified under the condition of exogeneity, OLS produced the best linear unbiased estimates, regardless of the distribution chosen for the error term. This finding is supported by the extant literature, as Gauss and Markov proved unbiasedness and efficiency of the OLS estimator without relying on assumptions of normality (Ebbes, 2004; Hueter, 2016). When endogenous data was generated from a linear regression model and the error term followed a uniform distribution, both IV and Gaussian copula methods produced unbiased, accurate results. The unbiased result found for the Gaussian copula approach matches the results reported by Park and Gupta (2012) for a uniform error distribution. However, once asymmetric, non-normal distributions were considered for the error term, the Gaussian copula no longer produced accurate results. The instrumental variable approach, given the true instrument, still produced highly accurate, unbiased results regardless of the distribution specified for the error term. These findings are new and suggest that the distributional assumption for the error term matters for the Gaussian copula approach, especially if the distribution is asymmetric.

For endogenous data generated from a LIV model where the error term was uniformly distributed, both the IV and LIV approach produced highly accurate, unbiased results. However, for asymmetric, non-normal distributions of the error, the LIV estimates became noticeably less accurate. Such results suggest that the LIV approach is somewhat sensitive to misspecification of the structural error term, which fits with findings reported by Ebbes (2009) in which the author noted that the LIV approach is more sensitive to distributional assumptions than other approaches, such as OLS. However, Ebbes' (2004) results for the LIV model when the error took on a Chi-square(1) distribution were found to be more accurate than those reported in this work when the error took on a Chi-square(4) distribution. While the two distributions have different parameters and are therefore different distributions, findings across these two different specifications were not expected to differ much. The author notes that the discrepancy in findings could very well be due to differences in the sampling methods used for the simulation, as Ebbes' samples from the two-stage equation error terms differently than done for this dissertation. However, further investigation is needed here. The IV regression approach, given the true, perfect instrument, was once again found to be robust to misspecification of the error term, producing highly accurate estimates regardless of the distribution specified.

The author subsequently investigated misspecifiation of the latent instrument, θ_i, and first-stage error term, v_i, in the two stage LIV equation. For misspecification, the author considered three distributions for both the latent instrument and the first-stage error: uniform; F-distribution(8,5); and Chi-square(4).

When the first-stage error term was uniformly distributed, both IV and LIV regression approaches produced unbiased, highly accurate results; however, when the distribution for the first-stage error became positively skewed (i.e., F- and Chi-square distributed), only IV regression produced highly accurate results, as LIV estimates became less accurate, albeit not statistically significantly so. Such results suggest that the distributional assumption placed on the first-stage error term is not overly restrictive. Although normality of the first-stage error term does not seem to be a strong assumption for the LIV approach, results suggest that symmetry of the first-stage error distribution matters. These findings are new, as misspecification of the first-stage error term in the LIV model had not previously been investigated to the author's knowledge.

Moreover, simulation results suggest that the LIV model preforms well when the latent instrument is misspecified. For all three distributions specified for the latent instrument, both the IV and the LIV model produced unbiased, accurate results. These findings are consistent with those reported by Ebbes (2009) and Papies et al. (2017), where the authors reported that the LIV model produced unbiased results when the latent instrument took on a non-normal gamma distribution. Findings from this dissertation work extend results from Ebbes (2009) and Papies et al. (2017) by considering an additional three distributions for the latent instrument that had not been previously explored. When viewed collectively, findings suggest that the assumption of a discrete multinomial distribution for the latent instrument is not a strong assumption of the LIV model, as the approach is robust to violations of this assumption.

Lastly, the author investigated misspecification of the first-stage error and latent instrument jointly. For this portion of the simulation study, the latent instrument was specified to follow a normal distribution and the first-stage error term followed a binary discrete distribution. As a result, the LIV model was severely misspecified for this research condition. Findings demonstrated that the LIV model produced highly inaccurate, biased results for this specification. However, the IV regression approach, provided with the true, perfect instrument, once again produced highly accurate, unbiased results.

Study 6. Simulation study six illustrated the importance of considering both instrument quality and sample size when using instrumental variable and instrument-free methods. For the instrumental variable approach, research findings demonstrated that the utility of the method for addressing endogeneity bias is largely dependent on the quality of the available instrument. When the available instrument is the true, perfect instrument, the IV regression method performs well, providing unbiased, consistent estimates; furthermore it outperformed the Gaussian copula approach in terms of mean-squared error (MSE) across all sample sizes investigated: *50*, *100, 250*, *500*, *1000*, *2500*, *and 5000.* Interestingly, the LIV regression approach performed equally well as IV regression using the true, perfect instrument across all investigated sample sizes. Moreover, research findings demonstrated that there was an efficiency cost for the IV approach when anything but the true, perfect instrument was used, even if the available instrument was of high quality and served as a good proxy for the true instrument. Once a proxy instrument was used for IV instead of the true instrument, the LIV modeling approach outperformed IV in terms of MSE at all investigated sample sizes.

Even when the available instrument was of high quality, IV regression still produced estimates farther on average from the true parameter value than OLS when endogeneity was minor and the sample size was insufficient. Specifically, for a simple linear regression specification with a single regressor and minor endogeneity bias, a sample size of 1,000 was required before IV regression outperformed OLS. This result supports findings from simulation research conducted by Boef, Dekkers, VandenBroucke, and le Cessie (2014), where the authors reported that IV often requires large sample sizes before outperforming OLS when endogeneity bias is minor.

Findings from this dissertation research also suggest that the degree of unmeasured confounding and sample size interact to determine the utility of instrument-free methods as well. For the Gaussian copula method, much larger sample sizes were needed to outperform OLS (N > 1,000), given minor endogeneity. For the LIV approach, simulation results suggested that only moderate sample sizes (N > 250) were required before outperforming OLS when there was minor endogeneity.

Further simulation research revealed that instrument-free methods performed better than instrumental variable regression across all investigated sample sizes once the available instrument became either weak, invalid or a combination of the two. Notably, when the available instrument was weak but exogenous, IV required very large sample sizes before providing reasonable estimates, as IV estimates demonstrated large bias and variance across most investigated samples. For the minor endogeneity condition, IV with a weak, exogenous instrument never outperformed OLS in terms of MSE; however, when endogeneity bias became severe, IV with a weak, exogenous instrument outperformed OLS at $N \geq 2{,}500$. These findings are consistent with the literature regarding IV regression with weak instruments (Bound et al., 1995; Crown et al., 2011; Boef et al., 2014; Hueter, 2016; Ebbes, 2004). Moreover, IV with a weak but exogenous instrument never outperformed instrument-free methods across sample sizes ranging from small to large.

Lastly, when the observed instrument became endogenous, IV regression underperformed both OLS and instrument-free approaches across all specified sample sizes and regressor-error endogeneity conditions, producing highly biased estimates with large variance. This finding is consistent with previous research regarding IV regression with bad quality instruments (Crown et al., 2011). Overall, simulation study six contributes new research findings to the field of causal inference, as to the author's knowledge there is no previous research considering the impact of sample size on instrument-free method performance.

Applied Findings: The academic impact of integrated student support

To investigate the causal impact of an integrated student support model, City Connects, on student academic achievement, the author applied 2SLS-IV and instrument-free regression methods to real-world school lottery data. Centralized assignment systems used by school districts to assign students to schools rely on random lotteries to break ties in admissions decisions. This creates a stratified RCT that researchers can take advantage of for conducting credible program evaluation research. Furthermore, by leveraging the random offer from lottery admissions, the researcher is afforded an arguably valid and high quality instrument for use in instrumental variable regression.

Results from Gaussian copula, LIV, and IV regression with a random lottery offer instrument demonstrated that the City Connects intervention had no impact on student academic achievement immediately post-randomization at kindergarten. The finding of non-significant differences was expected by the author, as given randomization, the students should be roughly equivalent on all covariates and outcome measures at this time point. Furthermore, City Connects is posited to be a long-term intervention that has impact over time by continually addressing students' strengths and needs (Chen, 2014; Lee-St. John, 2012). Interestingly, the LIV λ estimates, which reflect probability of group membership, were .53, reflecting nearly equal chances of group membership. This estimate accurately reflects the randomization process taking place at the kindergarten entry point and serves as further empirical validity evidence for the LIV modeling approach.

By the time students reached 3rd grade, both IV regression with a random offer lottery instrument and instrument-free methods revealed statistically significant positive achievement gains in mathematics for students receiving the City Connects intervention. Specifically, City Connects students received a predicted $.04\sigma$ to $.06\sigma$ increase in mathematics on a state test for every year the intervention was received relative to the counterfactual. Therefore, students randomized into the City Connects intervention at grade kindergarten and receiving the intervention through 3rd grade score $.16\sigma$ to $.24\sigma$ higher in mathematics on a state test than students not receiving the City Connects intervention during the same time period. This demonstrates a positive math effect for receiving the City Connects intervention between kindergarten and 3rd grade. By contrast, both IV and instrument-free methods suggested that City Connects had no significant impact on students' English Language Arts state test achievement at 3rd grade. For both mathematics and ELA, the estimates across methods were strikingly similar, and the IV and LIV math estimates matched nearly exactly. Such findings provide empirical validity evidence for the instrument-free methods.

Interestingly, both the IV and instrument-free method estimates were similar to the OLS estimate across analyses. The reason for this result is that endogeneity bias is likely not an issue for this sample. This claim was supported by endogeneity tests for all three methods. Specifically, the IV and Gaussian copula Hausman tests for endogeneity and the LIVρ estimate, a measure of endogeneity, all suggested exogeneity at the grade 3 time point. Lastly, the author estimated an observed LIV instrument from a nonparametric Bayesian LIV model and correlated this instrument with both the observed lottery offer instrument and the endogenous City Connects treatment variable. The LIV instrument was found to be highly correlated with the random lottery offer instrument (.74) and very highly correlated with the City Connects treatment variable (.99). Furthermore, the estimated LIV instrument correctly classified the student random offer for 86% of the observations. These results match earlier dissertation simulation findings, where the LIV instrument produced an estimated latent instrument that was strongly correlated with the true instrument. Overall, the findings from the lottery study offer new empirical validity evidence for the instrument-free methods, as, to the author's knowledge, this is the first time instrument-free methods have been applied to a large-scale real-world RCT. Moreover, this research presents new evidence regarding the efficacy of City Connects, as an analysis leveraging an RCT design for estimating the impact of City Connects has not been previously conducted. Dearing et al. (2016) revealed significant and practically important positive effects in mathematics performance during elementary school years for first-generation immigrant children living in high poverty, urban contexts and who received the City Connects intervention. Walsh et al. (2014) also reported higher mathematics performance for students participating in the City Connects intervention. However, such research findings are from quasi-experimental comparison group designs; the findings from this study are consistent with findings from the previous research investigating the impact of the City Connects intervention and provide stronger empirical evidence from a randomized control (RCT) trial design.

To make broader inferences about the impact of City Connects, instrument-free methods were then applied to the full district sample of students, regardless of whether or not they participated in the lottery assignment process. Results from both LIV and Gaussian copula regression approaches demonstrated that the City Connects intervention had no impact on student academic achievement at Kindergarten. The LIV λ estimates, which reflect probability of group membership, were .71 and .70, reflecting far from equal chances of group membership. This change in the estimate accurately reflects the full district data, as we are no longer taking advantage of a school lottery and therefore no longer have a RCT.

By the time students reached 3rd grade for the full district sample, instrument-free methods once again revealed statistically significant positive achievement gains in mathematics for students receiving the City Connects intervention. Specifically, for simple student-level models, City Connects students received a .03σ to .04σ increase in mathematics on a state test for every year the intervention was received relative to the counterfactual. When explicitly accounting for the multilevel structure of the data and between-school variability, the positive effect of receiving the City Connects intervention became even more notable. Specifically, for fixed effects and multilevel models, City Connects students received a .07σ to .08σ increase in mathematics on a state test for every year the intervention was received relative to the counterfactual. Therefore, students randomized into the City Connects intervention at grade kindergarten and receiving the intervention through 3rd grade score .28σ to .32σ higher in mathematics on a state test compared to students not receiving the City Connects intervention during the same time period. Surprisingly, both the Gaussian copula and LIV estimates were once again similar to the OLS estimate across analyses. Such a result suggests that endogeneity bias is likely not an issue even for the full district sample. This claim was supported by endogeneity tests for instrument-free methods. The Gaussian copula Hausman test for endogeneity and the LIVρ estimate, a measure of endogeneity, both suggested exogeneity.

Lastly, the researcher notes the lack of a statistically significant difference in reading and ELA achievement between City Connects and non-City Connects students across lottery and full sample research studies. It is speculated that this finding may be due to mathematics being primarily school-based and therefore its teaching and learning is much more confined to the contexts of classrooms than is the case with reading and language learning, as this type of learning can frequently take place outside of the school context. Consequently, school-based interventions may have more of an impact for mathematics achievement than they do for reading and ELA achievement; however, this topic merits further consideration and research.

Limitations and Future Research

In exploring the utility of the proposed methods, the author notes that there were several study limitations that warrant further research. Both limitations and future research will be discussed simultaneously in the sections that follow.

Estimation of Gaussian copula regression. Given Park and Gupta's (2012) formulation of the model, there are two ways one can estimate the Gaussian copula regression: least squares and maximum likelihood estimation (MLE). For this dissertation research, the author only considered estimation of the Gaussian copula regression using least squares. While Park and Gupta (2012) note that the Gaussian copula approach produced nearly identical results across estimation methods, this may not be the case for the research conditions investigated for this dissertation research. It is possible that one may get different results if a MLE Gaussian copula regression is used and it therefore may be worthwhile to investigate and compare a maximum likelihood based approach under the research conditions specified for this dissertation.

Dependence structure results. To the author's knowledge, this is the first time the dependence structure assumption for the proposed instrument-free methods has been discovered and explored. However, the reason for why the least squares Gaussian copula regression provides biased estimates for data generated from a LIV model including intercept but not for data generated from a LIV slope only specification remains unknown. Algebraically, the slope coefficient for a regression through the origin and a regression including intercept are nearly equivalent, differing only if the mean of the independent variable is not equal to zero (Kozak & Kozak, 1995). The researcher did investigate a mean-centering approach for the independent variable in further simulation study; however, the slope under the Gaussian copula approach remained biased. This result needs further investigation and possibly more detailed mathematical explanation as to why the inclusion of the intercept biases the Gaussian copula slope estimate under a LIV data generating process.

Furthermore, the LIV approach produces biased estimates when there exists no exogenous variation in the regressor and data is generated from a linear regression specification. Further investigation revealed that the estimated latent instrument only corrected for roughly half the endogeneity bias, and thus there was residual endogenous variation left over when a LIV modeling approach was fit to endogenous data generated from a linear regression specification. In other words, the LIV approach creates a less endogenous variable when the dependence structure is different from what is assumed for the model. To correct for this, one would only have to further adjust for the residual endogenous variation. Therefore, it may be possible to combine the two instrument-free approaches in such a way that the model appropriately accounts for any endogenous variation that could not be accounted for by the LIV modeling approach alone. Future research could explore a combined Gaussian copula LIV model, perhaps where the optimal LIV instrument is generated by the researcher and then combined with the Gaussian copula control function in some fashion. Furthermore, it may be possible in the LIV estimation process to account for any dependence between the latent instrument and structural error term.

Lastly, the researcher only identified that the dependence structure seems to matter for the instrument-free methods and, specifically, that the exogeneity requirement of the latent instrument breaks down if the dependence structure is different from what the LIV model assumes. It would be helpful if future research identified useful diagnostic checks or techniques for reliably identifying when the dependence structure differs from what each of the instrument-free methods assume.

Statistical models for sample size simulation. The sample size impact simulation study aimed to provide a starting point for generating useful guidelines and ideas about the appropriateness of the different methods under various contexts. However, the models investigated for the sample size simulation study were only simple linear regression models in that each method estimated an intercept and single slope only.

Such a model is often unrealistic in applied science, and furthermore by relying on these simple models for providing sample size guidelines one may underestimate the actual sample size needed by applied researchers. This is because the models often employed in applied research have many more parameters to estimate, which ties up additional degrees of freedom and thus require larger samples. As a result, the impact of sample size on instrument-free methods and instrumental variable regression should be further investigated using more fully parametrized models.

In addition to the consideration of more fully parametrized models, one may also wish to consider various treatment assignment mechanisms and degrees to which the instrumental variable influences selection when investigating the impact of sample size on IV and instrument-free methods. Research conducted by Boef et al. (2014) may serve as a useful starting point for performing this comparison. Furthermore, it would be interesting to see how well the equation Boef et al. (2014) derived for approximating the threshold sample size at which IV outperforms OLS also approximates the sample sizes needed for instrument-free methods to outperform OLS. Lastly, it should be theoretically possible to derive an equation for approximating the threshold sample size at which instrument-free methods outperform OLS in terms of MSE.

Consistency of the LIV estimator. This dissertation research contributes to the prior research conducted by Ebbes (2004; 2005; 2009), Papies et al. (2017), and Park and Gupta (2012) by also demonstrating that the LIV estimator is approximately consistent via simulation study results. However, we reserve the term ‘approximately consistent’ for describing the LIV estimator due to the fact that consistency has only been shown via simulation study. To the author’s knowledge, a mathematical proof for the consistency of the LIV estimator still does not exist. Future mathematical research proving this property would ensure confidence in applying the LIV modeling approach.

Coding of the lottery random offer instrument. While the random offer instrument from an admissions lottery is a valid instrument, coding decisions that affect the representational accuracy of the instrument were made. The actual random offer provided in the admissions data was for specific schools. The researcher then coded this random offer to a specific school as being a random offer to a City Connects school if the school the student received an offer to attend was ever a City Connects school across a number of years. In other words, an if-ever coding scheme was used. This approach is not representative of actual reality and possibly weakens the instrument; however, it should not affect exogeneity of the instrument.

Level of City Connects treatment. Given the individualized nature of the City Connects intervention, the unit of randomization being the student for the lottery study, and the dosage measure of exposure to treatment, the City Connects treatment was specified to be at the student-level for all applied analyses. However, it is important to note that City Connects is a school-wide intervention, and furthermore that treatment may be heterogeneous at the school-level. Therefore it is possible that treatment and the models used for the applied analyses were misspecified. As a possible solution, one may wish to consider modeling treatment at both the student- and school-level in future analyses. Additionally, although the lottery study explored for this dissertation research randomizes students to schools, it is important to note that schools have not been randomized to the City Connects treatment condition. It is therefore also possible that important school-level factors, e.g., effective principal leadership, account for the treatment effects observed.

Treatment dosage. By including years spent in the City Connects intervention as a measure of dosage in our models, we are assuming that there is a linear increase in outcome performance for each year spent in the City Connects intervention. This is a strong assumption, and one which may be unrealistic. Future research should empirically investigate the degree to which this assumption holds by including nonlinear transformations of the treatment variable in regression models.

Broader Implications

Instrument-free methods are useful approaches for addressing endogeneity bias and providing researchers with valuable information about the causal impact of an intervention. Given the problem of finding high quality instruments, and the potentially even bigger problem of relying on poor quality instruments, it is important to have alternative statistical approaches for addressing unobserved confounding when probing causal hypotheses. Instrument-free statistical approaches serve as viable alternatives to instrumental variable regression, providing unbiased, accurate causal estimates across a range of research conditions involving endogeneity bias. Furthermore, these methods do not require that the researcher identify a valid instrument. However, as research has also demonstrated, these approaches require their own set of identifying assumptions before one can infer causality (Papies et al., 2017). This underscores the importance of assumptions when performing causal inference in the absence of experimental data. Because all non-experimental research relies on key assumptions, no one method in statistical causal inference is invariably superior to another, and therefore we must be aware of the assumptions made under any given modeling approach and the degree to which these assumptions fit our context (Papies et al., 2017).

When performing causal inference with OLS, we make the key assumption that the treatment variable is uncorrelated with the structural error term. For IV regression, the assumptions change, and we now assume that an instrumental variable is available for use and that this variable is both relevant and exogenous. Such assumptions may be hard to satisfy, and this strongly encourages the use of an instrument-free approach. However, in adopting an instrument-free modeling approach, we must be aware that we are making new and important assumptions. Specifically, the assumptions of instrument relevance and exogeneity are replaced by the assumptions that the endogenous regressor is non-normal and that there exists a certain dependence structure in place. For the LIV model, we assume additive separability in the endogenous regressor, and that the dependence can be captured by the correlation between the structural error term and the additive endogenous component of the endogenous regressor. For the Gaussian copula approach, we assume that the exogeneity requirement does not hold, and that the dependence between the endogenous regressor and the structural error term can be represented as purely correlational. In sum, we use different sets of assumptions for making causal inferences under different approaches, and the validity of the causal inference is tied to the degree to which assumptions hold for any given analytic approach.

This dissertation research also offers new empirical validity evidence for instrument-free methods by applying the approaches to a real-world large-scale RCT. As demonstrated, the results from instrument-free methods matched results from IV regression using a random lottery offer instrument from an RCT. Moreover, the latent instrument produced by the LIV approach strongly correlated with the observed random lottery offer instrument. Such results are very encouraging and demonstrate the power of instrument-free methods. Given this result, educational researchers should consider the adoption of instrument-free methods in their own substantive work. To the author's knowledge, instrument-free methods have not yet been adopted for educational evaluation research purposes. Hopefully, this practice will change as these methods align well with education researchers' substantive goals and could be a powerful complement to other more commonly employed regression techniques. This is especially the case for education researchers seeking to make causal claims about student services and academic interventions.

In line with the above, instrument-free methods were used in combination with IV regression to demonstrate the academic impact of an integrated student support model, namely the City Connects intervention. Findings revealed significant positive effects for students receiving the intervention during early elementary school years. Such findings are consistent with the extant research examining the efficacy of the City Connects student support model. Dearing et al. (2016) noted a significant and practically important positive effect in both mathematics and reading performance for first-generation immigrant children in high poverty, urban contexts during elementary school years associated with exposure to the City Connects intervention. Furthermore, Walsh et al. (2014) reported both higher report card scores and higher performance on middle school English language arts and mathematics tests for students participating in the City Connects intervention. These studies provide compelling evidence for the effectiveness of City Connects for addressing non-academic barriers to learning. This dissertation research contributes to these findings by offering efficacy evidence from a real-world RCT. Additionally, instrument-free methods were used to triangulate findings and further support claims regarding the impact of the City Connects intervention.

The Every Student Succeeds Act (ESSA), the nation's main law for public education, has shifted focus toward more disadvantaged students by encouraging the use of integrated student support programs for addressing barriers to learning brought on by poverty and other contextual factors (ESSA Title I, Title IVA). Stemming from this federal law, numerous states have adopted legislation to advance integrated student support strategies (Policy Brief, p.14-17). City Connects is an evidence-based student support model that demonstrates the feasibility of offsetting the impact of out-of-school factors on learning and healthy development through the provision of comprehensive, tailored and individualized services (Dearing et al., 2016; Walsh et al., 2014; Progress Report, 2016). As a result, the City Connects intervention can inform the development of state policies and protocols, and guide more effective approaches to implementation at scale.

REFERENCES

Abdulkadiroglu, A., Pathak, P., Roth, A., & Sonmez, T. (2006). Changing the boston school choice mechanism: Strategy-proofness as equal access. MIT Files. Retrieved from: https://economics.mit.edu/files/3030

Abdulkadiroglu, A., Angirst, J., Dynarski, S., Kane, T., & Pathak, P. (2011). Accountability and flexibility in public schools: Evidence from boston's charters and pilots. *The Quarterly Journal of Economics, Volume*(126), 699-748. The Oxford Press. doi:10.1093/qje/qjr017

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., & Pathak, P. A. (2017). Research design meets market design: Using centralized assignment for impact evaluation. *Econometrica*, *85*(5), 1373-1432.

Adelson, J. (2013). Educational research with real-world data: Reducing selection bias with propensity scores.

Alemayehu, D., Alvir, J.M., Jones, B., & Willke R.J. (2011). Statistical issues with the analysis of nonrandomized studies in comparative effectiveness research. *Journal of Managed Care Pharmacy, Volume*(9). doi: 10.18553/jmcp.2011.17.s9-a.S22

Angrist, J. & Krueger, A. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics, Volume*(106), 979-1014.

Berliner, D. (2009). *Poverty and Potential: Out-of-School Factors and School Success*. Boulder and Tempe: Education and the Public Interest Center & Education Policy

Research Unit. Retrieved [7/1/2016] from:
http://files.eric.ed.gov/fulltext/ED507359.pdf

Boef, A., Dekkers, O., Vandenbroucke, J., & le Cessie, S. (2014). Sample size importantly limits the usefulness of instrumental variables methods, depending on instrument strength and level of confounding. *Journal of Clinical Epidemiology,* (67),

Bound, J., Jaeger, D., & Baker, R. (1995). Problems with instrumental variables estimation when correlation between the instrument and the endogenous explanatory variable is weak. *Journal of the American Statistical Association, Volume*(90), 443-450.

Brooks, S., Gelman, A., Jones, G., & Meng, X. L. (Eds.). (2011). *Handbook of markov chain monte carlo*. CRC press.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297-312.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. *Handbook of research on teaching. Chicago, IL: Rand McNally*.

Center for Promise & Center for Optimized Student Support (2019). Building systems of integrated student support: A policy brief for local and state leaders. Retrieved from: https://www.americaspromise.org/sites/default/files/d8/ISSreport_v7_0.pdf

Chen, J. (2014). *A Bayesian Propensity Score Approach for Multilevel Observational Studies* (Doctoral dissertation).

Chesher, A. & Rosen, A. (2015). What do instrumental variable methods deliver with discrete dependent variables? (Working Paper). The Institute for Fiscal Studies, Department of Economics, UCL. Retrieved from: https://www.cemmap.ac.uk/wps/cwp101313.pdf

City Connects (2016). The impact of City Connects: Student outcomes. Progress report 2016. Retrieved from: https://www.bc.edu/content/dam/files/schools/lsoe/cityconnects/pdf/City%20Connects%20Progress%20Report%202016.pdf

City Connects (2018). City Connects: Intervention and impact. Progress report 2018. Retrieved from: https://www.bc.edu/content/dam/bc1/schools/lsoe/sites/coss/pdfs/CityConnectsProgressReport2018.pdf

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Crown, W.H., Henk, H.J., & Vanness, D.J. (2011). Some cautions on the use of instrumental variables estimators in outcomes research: How bias in instrumental variables estimators is affected by instrument strength, instrument contamination, and sample size. *Value Health, Volume*(8), 1076-1084. doi: 10.1016/j.jval.2011.06.009.

Dearing, E., Walsh, M. E., Sibley, E., Lee-St. John, T., Foley, C., & Raczek, A. E. (2016). Can Community and School-Based Supports Improve the Achievement of

First-Generation Immigrant Children Attending High-Poverty Schools?. *Child development*, *87*(3), 883-897.

Deaton, A., & Cartwright, N. (2017). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*.

Do, C. (2008). The multivariate Gaussian distribution [CS229 Lecture Notes]. Retrieved from: http://cs229.stanford.edu/section/gaussians.pdf

Duncan, G. J., Yeung, W. J., Brooks-Gunn, J., & Smith, J. R. (1998). How much does childhood poverty affect the life chances of children?. *American sociological review*, *63*(3), 406.

Ebbes, P. (2004). Latent instrumental variables: A new approach to solve for endogeneity. Thesis. Retrieved from: https://www.rug.nl/research/portal/files/13172416/thesis.pdf

Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, T. (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics*, *3*(4), 365-392.

Ebbes, P., Wedel, M., & Böckenholt, U. (2009). Frugal IV alternatives to identify the parameter for an endogenous regressor. *Journal of Applied Econometrics*, *24*(3), 446-468.

Ebbes, P., Bockenholt, U., Wedel, M., & Nam, H. (2014). Accounting for regressor-error dependencies in educational data: A bayesian mixture approach (Working Paper ID 2466533).

Embrechts, P., Lindskog, F., & McNeil, A. (2001). Modelling Dependence with Copulas and Applications to Risk Management. Retrieved from: https://people.math.ethz.ch/~embrecht/ftp/copchapter.pdf

Every Student Succeeds Act of 2015, Pub. L. 114-95.

Faye, J. (2014). Causal explanations in *The Nature of Scientific Thinking.* Palgrave Macmillan, London.

Freedman, D. (1997). From association to causation via regression. *Advances in applied mathematics*, *18*(1), 59-110.

Freedman, D. A. (2005). Linear statistical models for causation: A critical review. *Wiley StatsRef: Statistics Reference Online*.

Gates, E., & Dyson, L. (2017). Implications of the Changing Conversation About Causality for Evaluators. *American Journal of Evaluation*, *38*(1), 29-46.

Gerring, J. (2011). Social Science Methodology: A Unified Approach. Cambridge University Press.

Glynn, A., & Quinn, K. (2007). *Non-parametric Mechanisms and Causal Modeling*. working paper.

Görür, D., & Rasmussen, C. E. (2010). Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, *25*(4), 653-664.

Gui, R., Meierer, M., Algesheimer, R., & Schilter, P. (2019). REndo: Fitting Linear Mod els with Endogenous Regressors using Latent Instrumental Variables. R packag e Version 2.2.0. https://CRAN.R-project.org/package=REndo

Hastings, J., Kane, T., & Staiger, D. (2009). Heterogenous preferences and the efficacy of public school choice. Retrieved from: http://justinehastings.com/wp-content/uploads/2016/12/HKS_Combined_201006.pdf

Hernan, M. & Robins, J. (2013). *Causal Inference*. Taylor & Francis

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*(396), 945-960. doi:10.2307/2289064

Holland, P. W., & Rubin, D. B. (1987). Causal inference in retrospective studies. *ETS Research Report Series*, *1987*(1), 203-231.

Hueter, I. (2016). Latent instrumental variables: A critical review (Working Paper No. 46) New York, NY: Institute for New Economic Thinking.

Hulswit, M. (2002). *From cause to Causation: A Peircean Perspective.* Kluwer Academic Publishers. Hume, D. (1748). *Philosophical Essays Concerning Human Understanding: By the Author of the Essays Moral and Political*. A. Millar.

Huurre, T., Aro, H., Rahkonen, O., & Komulainen, E. (2006). Health, lifestyle, family and school factors in adolescence: predicting adult educational level. *Educational research*, *48*(1), 41-53.

Kaplan, D. (2009). Causal inference in non-experimental educational policy research in *Handbook of Education Policy Research.* New York, NY: Routledge.

Kaushal, N., Magnuson, K., & Waldfogel, J. (2012). How is family income related to investments in children's learning?.

Keppel, G. & Wickenss, T. (2004). *Design and Analysis: A Researcher's Handbook* (4th ed.). Prentice Hall.

Kleiber, C. & Zeileis, A. (2008). Applied Econometrics with R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. https://CRAN.R-project.org/package=AER

Lee-St. John, T.J. (2012). The causal impact of systematic, elementary school student support on middle school grade retention. (Dissertation).

Lee-St. John, T. J., Walsh, M. E., Raczek, A. E., Vuilleumier, C. E., Foley, C., Heberle, A., ... & Dearing, E. (2018). The Long-Term Impact of Systemic Student Support in Elementary School: Reducing High School Dropout. *AERA Open*, *4*(4), 2332858418799085

Lunn, D., Jackson, C., Best, N., Spiegelhalter, D., & Thomas, A. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Chapman and Hall/CRC.

O'Dwyer, L. & Parker, C. (2014). A primer for analyzing nested data: multilevel modeling in SPSS using an example from a REL study. Retrieved from: http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2015046.pdf

Ozen, U. (2011). Causality in philosophy in science: From hume to today. *International Journal of Humanities and Social Science, 1*(11). Retrieved from: http://www.ijhssnet.com/journals/Vol_1_No_11_Special_Issue_August_2011/22.pdf

Papies, D., Ebbes, P., & Van Heerde, H. J. (2017). Addressing endogeneity in marketing models. In *Advanced methods for modeling markets* (pp. 581-627). Springer, Cham.Park, S. & Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science, 31*(4), 567-586.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Survey, Volume*(3), 96-146. DOI: 10.1214/09-SS057

Peterson, J. (1898). The empirical theory of causation. *The Philosophical Review, 7*(1). Retrieved from: https://www.jstor.org/stable/2175547?Search=yes&resultItemClick=true&searchText=peterson&searchText=causal&searchText=inference&searchUri=%2Faction%2FdoBasicResults%3Fsi%3D1%26amp%3Bgroup%3Dnone%26amp%3Bso%3Dold%26amp%3BQuery%3Dpeterson%2Bcausal%2Binference%26amp%3Brefreqid%3Dsearch%253A93abb8f77df55de1c80f1fbc4d0e871b%26amp%3Bfc%3Doff%26amp%3Bacc%3Don%26amp%3Bwc%3Don&refreqid=search%3Ae699e271269fe02fbcde1246515a01bc&seq=1#page_scan_tab_contents

Rodriguez, G. (2007). Lecture Notes. Retrieved from: http://data.princeton.edu/wws509/notes/c2.pdf

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175-214.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Shadish, W., Cook, T., & Campbell, D. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Miflin

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological methods*, *15*(1), 3.

Shalizi, C. (2015). Mixture Models (Lecture Notes). Retrieved from: http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch20.pdf

Shields, K.A., Walsh, M.E., & Lee- St. John, T.J. (2016). The Relationship of a Systematic Student Support Intervention to Academic Achievement in Urban Catholic Schools. *Journal of Catholic Education, 19* (3). http://dx.doi.org/10.15365/joce.1903072016

Steele, J. L., Slater, R. O., Zamarro, G., Miller, T., Li, J., Burkhauser, S., & Bacon, M. (2017). Effects of dual-language immersion programs on student achievement: Evidence from lottery data. *American Educational Research Journal*, *54*(1_suppl), 282S-306S.

Stock, J. & Watson, M. (2014). *Introduction to Econometrics*. Pearson Series.

Stuart Mill, J. (1843). A system of logic. *Ratiocinative and Inductive, Cosimo, New York, NY.*

Tilly, J. & Janetos, N. (2018). matchingR: Matching Algorithms in R and C++. R packag e version 1.3.0. https://CRAN.R-project.org/package=matchingR

Van De Geer, S. (2005). Least Squares Estimation. *Encyclopedia of Statistics in* Behavioral *Science.* John Wiley & Sons.

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, *15*(1), 18.

Willett, J. & Murnane, R. (2010). *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* Oxford University press.

Whye The, Y. (2010). Dirichlet Process. Retrieved from: https://www.stats.ox.ac.uk/~teh/research/npbayes/Teh2010a.pdf

Walsh, M. (2013). Support All Students to Close the Achievement Gap. The Whole Child Blog. Retrieved from: http://www.wholechildeducation.org/blog/support-all-students-to-close-the-achievement-gap

Walsh, M. E., Madaus, G. F., Raczek, A. E., Dearing, E., Foley, C., An, C., ... & Beaton, A. (2014). A new model for student support in high-poverty urban elementary schools: Effects on elementary and middle school academic outcomes. *American Educational Research Journal*, *51*(4), 704-737.

Walsh, M., Raczek, A., Sibley, E., Lee- St. John, T., An, C., Akbayin, B., Dearing, E., & Foley, C. (2015). City Connects: Building an Argument for Effects on Student

Achievement with a Quasi-Experimental Design (Conference Abstract). Retrieved from: http://files.eric.ed.gov/fulltext/ED562168.pdf

APPENDIX

Table 1.

Impact of City Connects with Dichotomous Dose variable

		OLS		IV	
Grade	*Subject*	$\hat{\beta}(S.E.)$	*p*-value	$\hat{\beta}(S.E.)$	*p*-value
K	Math	.06 (.04)	0.13	.18 (.15)	0.25
	Reading	-.02 (.04)	0.62	.04 (.14)	0.76

Table 2.

MC Error and $\hat{R}$ statistics for kindergarten ELA randomization sample

Parameter	**S.D.**	**MC Error**	$\hat{R}$
Intercept	0.072	0.004	1.00
City Connects	0.006	< .001	1.00
Year 07	0.062	0.002	1.00
Year 08	0.070	0.002	1.00
Year 09	0.066	0.002	1.00
Year 10	0.065	0.003	1.00
Year 11	0.066	0.002	1.00
Year 12	0.061	0.002	1.00
Year 13	0.081	0.003	1.00
Male	0.034	<.001	1.00
Black	0.071	0.004	1.00
Hispanic	0.068	0.004	1.00
Asian	0.080	0.004	1.00
Mixed	0.130	0.005	1.00
Special Ed. 1	0.140	0.004	1.00
Special Ed. 2	0.085	0.003	1.00
Special Ed. 3	0.242	0.006	1.00
Free Lunch	0.050	0.002	1.00
Reduced Lunch	0.096	0.003	1.00
ELL	0.044	0.001	1.00
λ_1	0.010	< .001	1.00
λ_2	0.010	< .001	1.00
π_1	0.008	< .001	1.00
π_2	0.008	< .001	1.00
ρ	0.058	0.002	1.00
σ_1	0.021	< .001	1.00
σ_2	0.002	< .001	1.00

Table 3.

MC Error and $\hat{R}$ statistics for grade 3 ELA randomization sample

Parameter	**S.D.**	**MC Error**	$\hat{R}$
Intercept	0.094	0.005	1.00
City Connects	0.017	< .001	1.00
Year 10	0.078	0.003	1.01
Year 11	0.079	0.003	1.00
Year 12	0.076	0.003	1.00
Year 13	0.077	0.002	1.00
Male	0.047	0.001	1.00
Black	0.094	0.004	1.01
Hispanic	0.087	0.004	1.00
Asian	0.106	0.004	1.00
Mixed	0.182	0.006	1.00
Special Ed. 1	0.159	0.004	1.00
Special Ed. 2	0.107	0.003	1.01
Special Ed. 3	0.196	0.006	1.00
Free Lunch	0.076	0.003	1.00
Reduced Lunch	0.132	0.004	1.00
ELL	0.068	0.002	1.00
λ_1	0.014	< .001	1.00
λ_2	0.014	< .001	1.00
π_1	0.022	< .001	1.00
π_2	0.031	0.001	1.00
ρ	0.071	0.002	1.00
σ_1	0.029	< .001	1.00
σ_2	0.012	< .001	1.00

Table 4.

MC Error and $\hat{R}$ statistics for kindergarten ELA full sample

Parameter	**S.D.**	**MC Error**	$\hat{R}$
Intercept	0.022	< .001	1.00
City Connects	0.002	< .001	1.00
Year 05	0.022	< .001	1.00
Year 06	0.022	< .001	1.00
Year 07	0.022	< .001	1.00
Year 08	0.022	< .001	1.00
Year 09	0.022	< .001	1.00
Year 10	0.022	< .001	1.00
Year 11	0.021	< .001	1.00
Year 12	0.022	< .001	1.00
Year 13	0.024	< .001	1.00
Male	0.010	< .001	1.00
Black	0.017	< .001	1.00
Hispanic	0.017	< .001	1.00
Asian	0.023	< .001	1.00
Mixed	0.035	< .001	1.00
Special Ed. 1	0.044	0.001	1.00
Special Ed. 2	0.020	< .001	1.00
Special Ed. 3	0.058	0.002	1.00
Free Lunch	0.015	< .001	1.00
Reduced Lunch	0.028	< .001	1.00
ELL	0.013	< .001	1.00
λ_1	0.002	< .001	1.00
λ_2	0.002	< .001	1.00
π_1	0.003	< .001	1.00
π_2	0.004	< .001	1.00
ρ	0.012	< .001	1.00
σ_1	0.007	< .001	1.00
σ_2	0.001	< .001	1.00

Table 5.

MC Error and $\hat{R}$ statistics for grade 3 ELA full sample

Parameter	**S.D.**	**MC Error**	$\hat{R}$
Intercept	0.021	< .001	1.00
City Connects	0.005	< .001	1.00
Year 02	0.020	< .001	1.01
Year 03	0.021	< .001	1.00
Year 04	0.021	< .001	1.01
Year 05	0.021	< .001	1.00
Year 06	0.021	< .001	1.00
Year 07	0.021	< .001	1.00
Year 08	0.022	< .001	1.00
Year 09	0.011	< .001	1.00
Year 10	0.022	< .001	1.00
Year 11	0.021	< .001	1.00
Year 12	0.021	< .001	1.00
Year 13	0.022	< .001	1.00
Male	0.009	< .001	1.00
Black	0.014	< .001	1.00
Hispanic	0.014	< .001	1.01
Asian	0.020	< .001	1.01
Mixed	0.040	0.001	1.00
Special Ed. 1	0.040	0.001	1.00
Special Ed. 2	0.017	< .001	1.00
Special Ed. 3	0.024	< .001	1.00
Free Lunch	0.016	< .001	1.00
Reduced Lunch	0.027	< .001	1.01
ELL	0.011	< .001	1.00
λ_1	0.001	< .001	1.00
λ_2	0.001	< .001	1.00
π_1	0.002	< .001	1.00
π_2	0.006	< .001	1.00
ρ	0.012	< .001	1.00
σ_1	0.006	< .001	1.00
σ_2	<.001	< .001	1.00

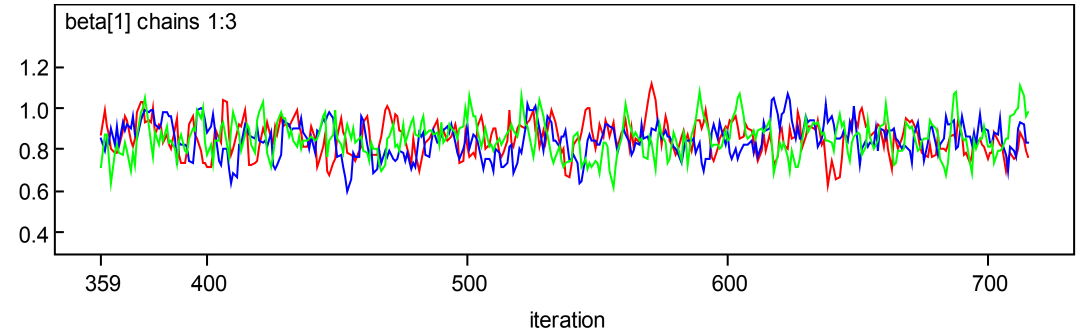

Figure 1. Trace Plot for Kindergarten Lottery Sample Math Intercept

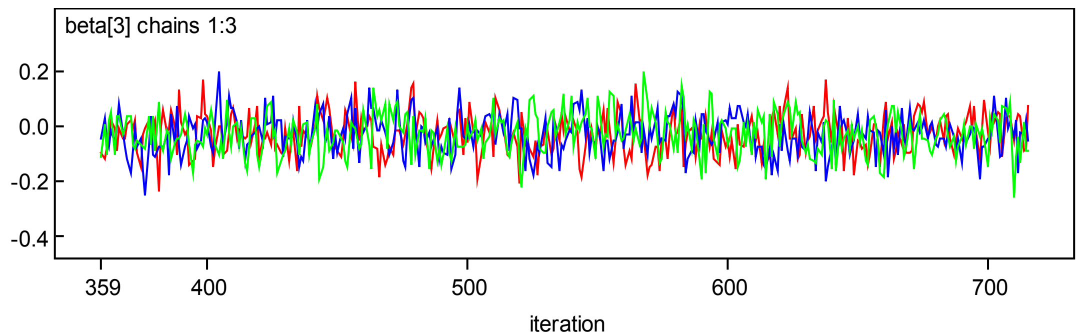

Figure 2. Trace Plot for Kindergarten Lottery Sample Math Year 07

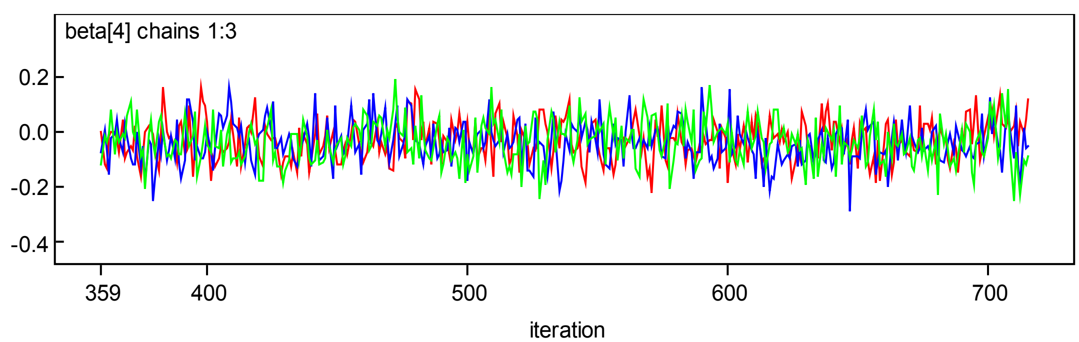

Figure 3. Trace Plot for Kindergarten Lottery Sample Math Year 08

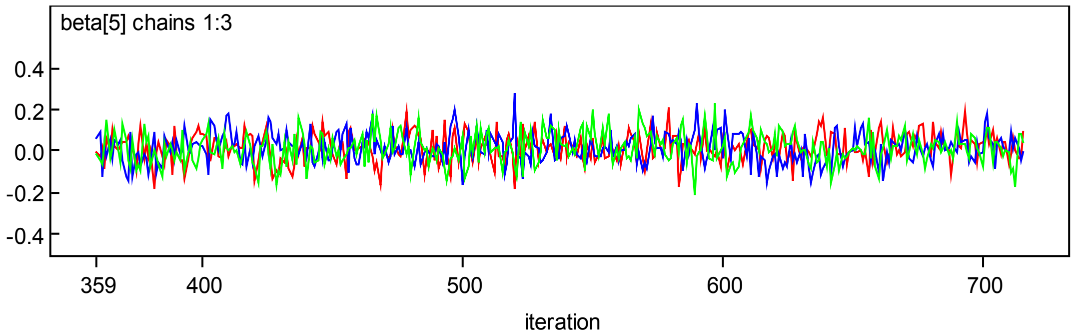

Figure 4. Trace Plot for Kindergarten Lottery Sample Math Year 09

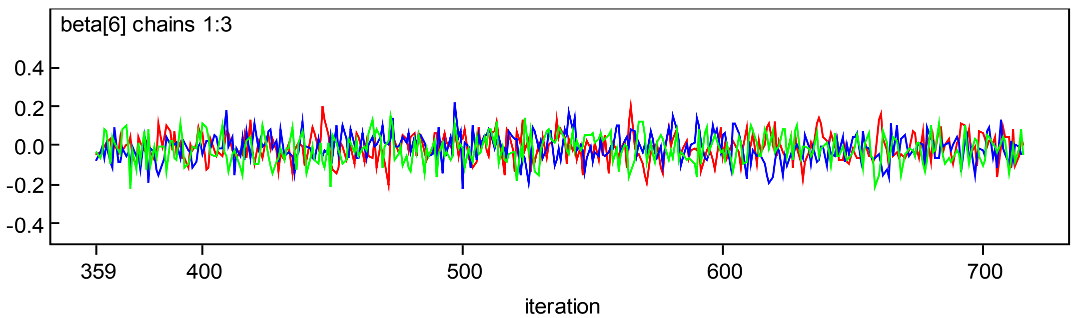

Figure 5. Trace Plot for Kindergarten Lottery Sample Math Year 10

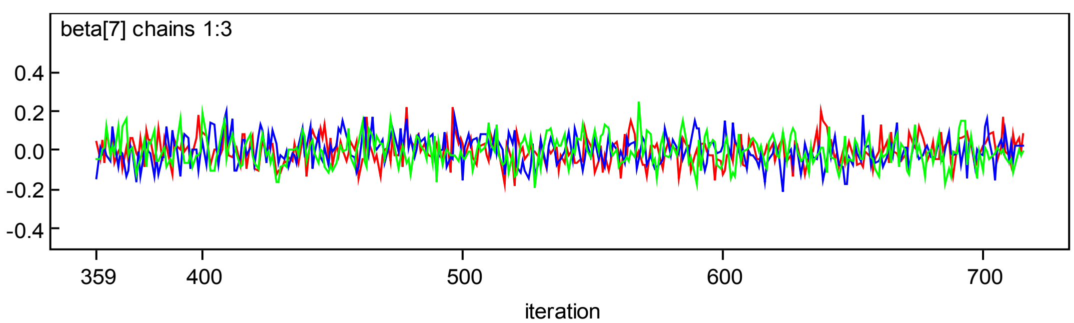

Figure 6. Trace Plot for Kindergarten Lottery Sample Math Year 11

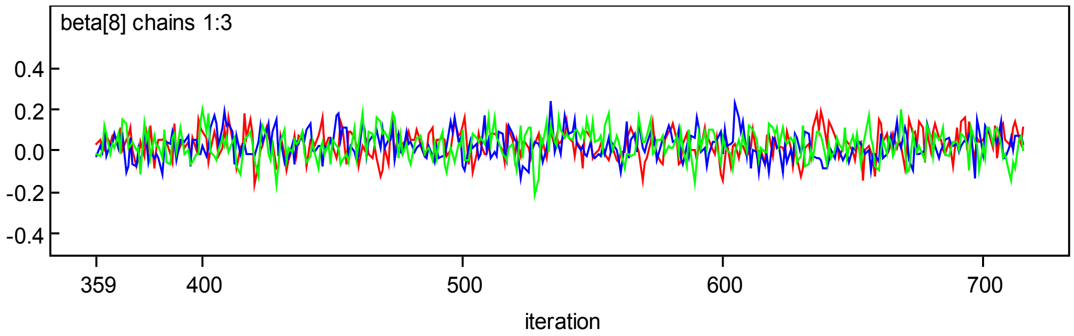

Figure 6. Trace Plot for Kindergarten Lottery Sample Math Year 12

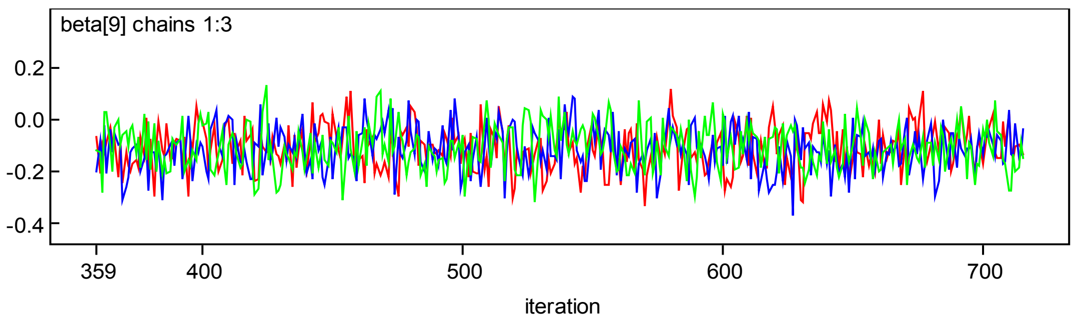

Figure 7. Trace Plot for Kindergarten Lottery Sample Math Year 13

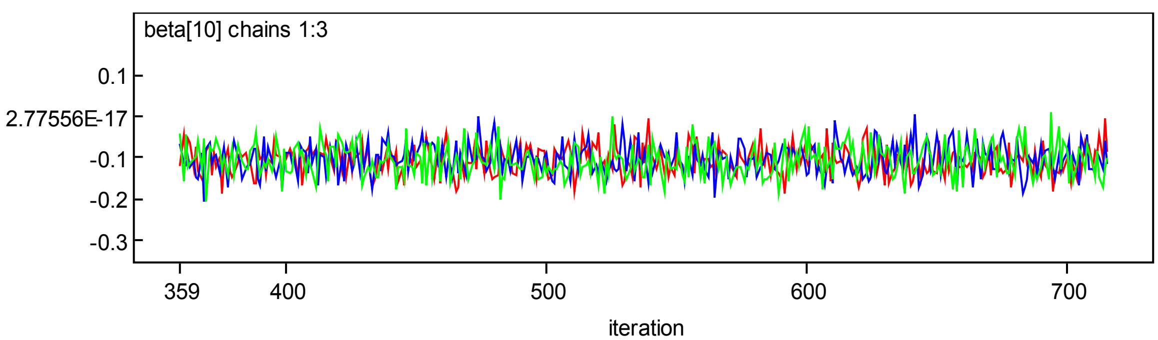

Figure 8. Trace Plot for Kindergarten Lottery Sample Math Male

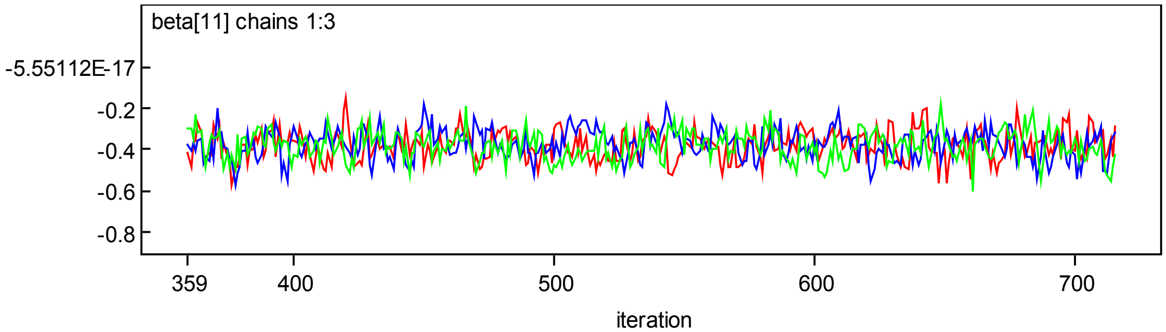

Figure 9. Trace Plot for Kindergarten Lottery Sample Math Black

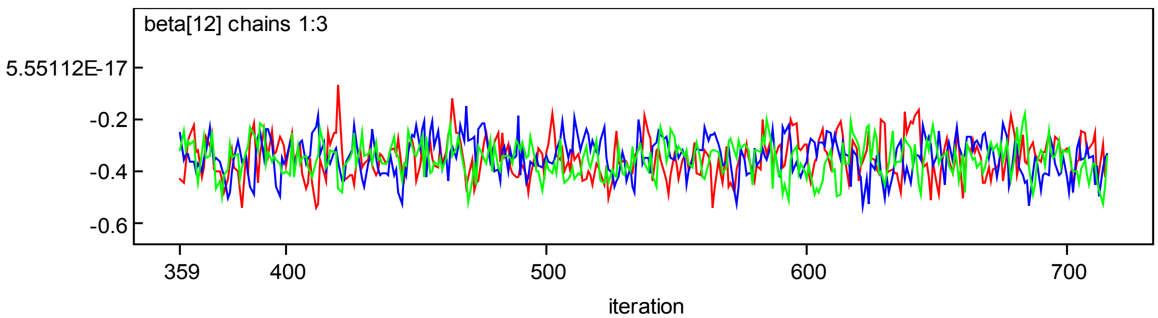

Figure 10. Trace Plot for Kindergarten Lottery Sample Math Hispanic

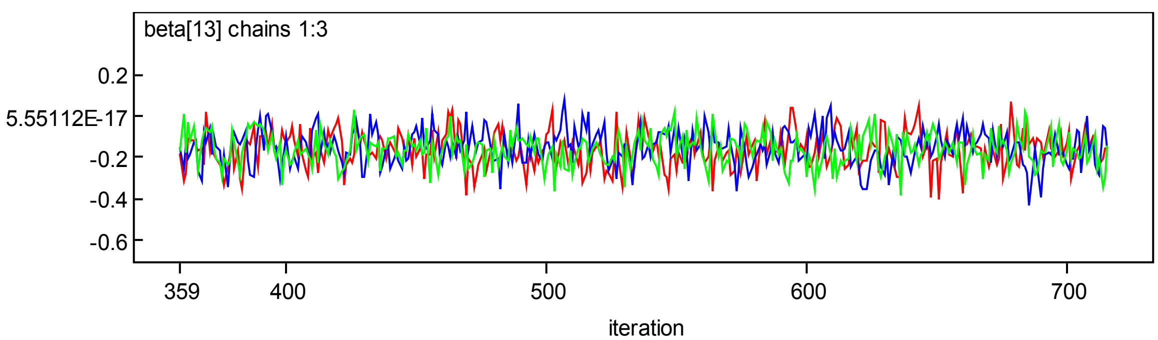

Figure 11. Trace Plot for Kindergarten Lottery Sample Math Asian

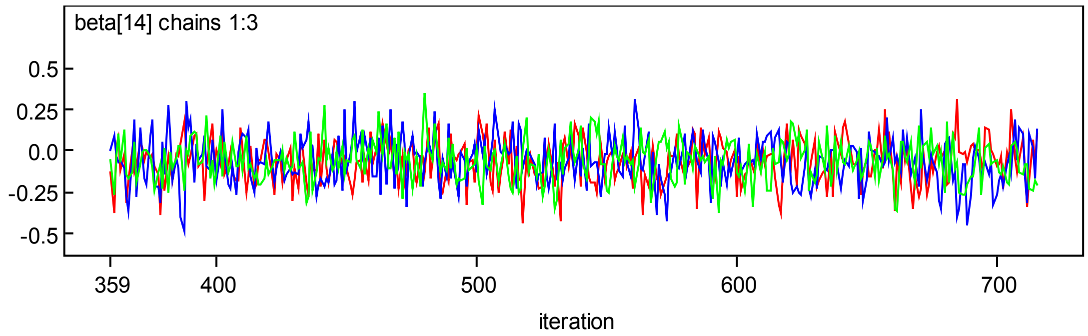

Figure 12. Trace Plot for Kindergarten Lottery Sample Math Mixed

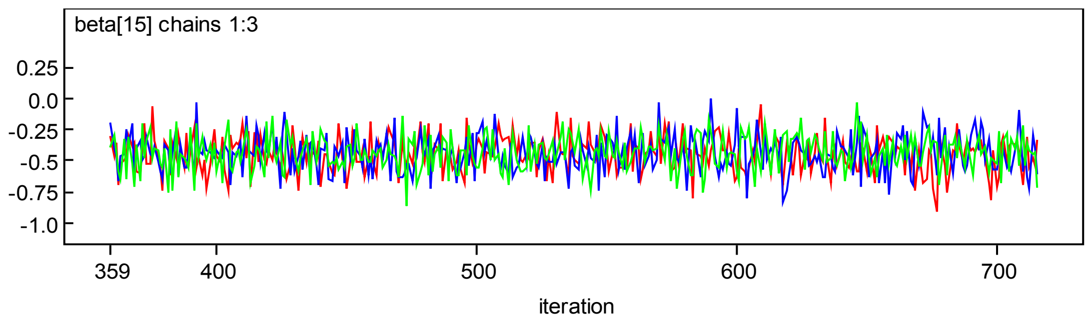

Figure 13. Trace Plot for Kindergarten Lottery Sample Math Sped 1

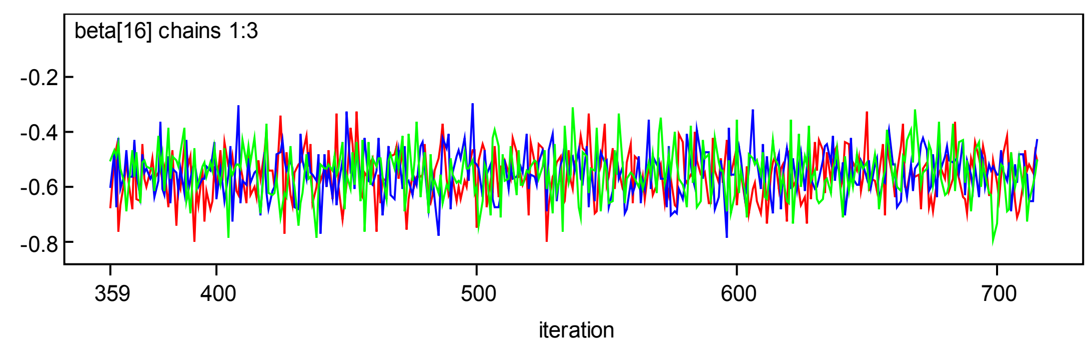

Figure 14. Trace Plot for Kindergarten Lottery Sample Math Sped 2

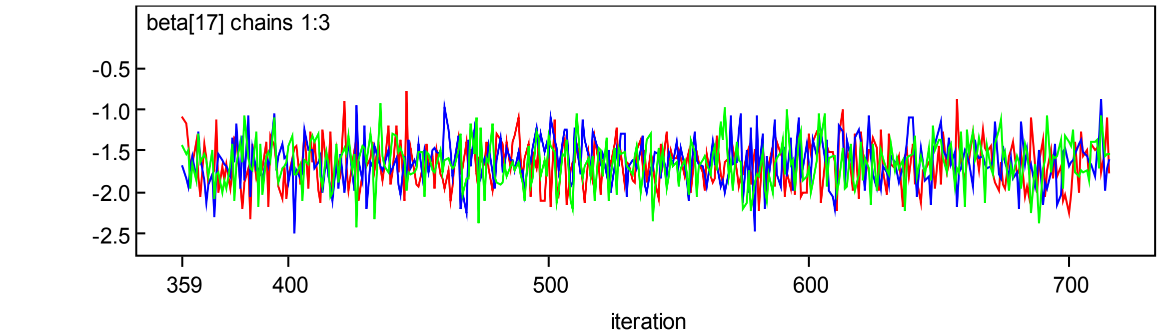

Figure 15. Trace Plot for Kindergarten Lottery Sample Math Sped 3

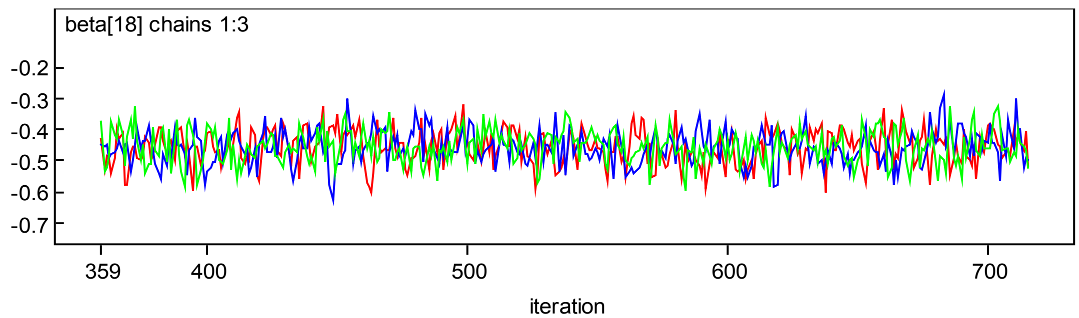

Figure 16. Trace Plot for Kindergarten Lottery Sample Math Free Lunch

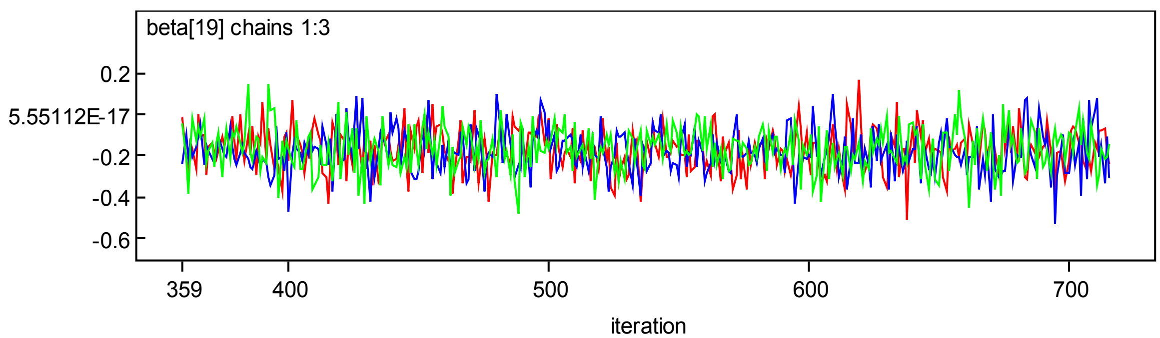

Figure 17. Trace Plot for Kindergarten Lottery Sample Math Reduced Lunch

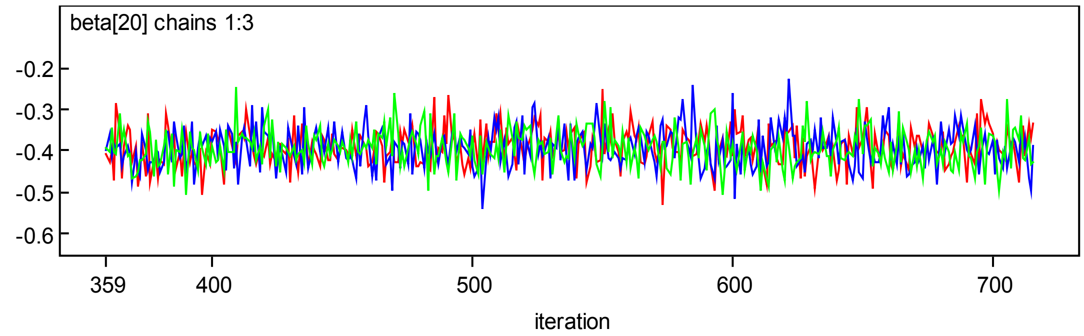

Figure 18. Trace Plot for Kindergarten Lottery Sample Math Bilingual

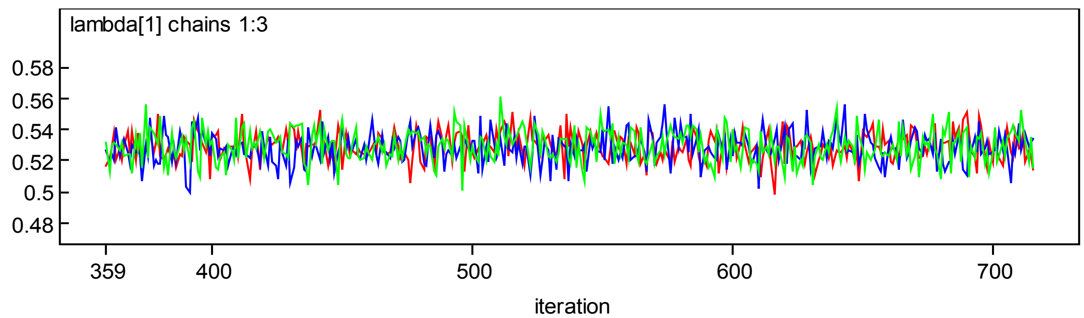

Figure 19. Trace Plot for Kindergarten Lottery Sample Math Lambda 1

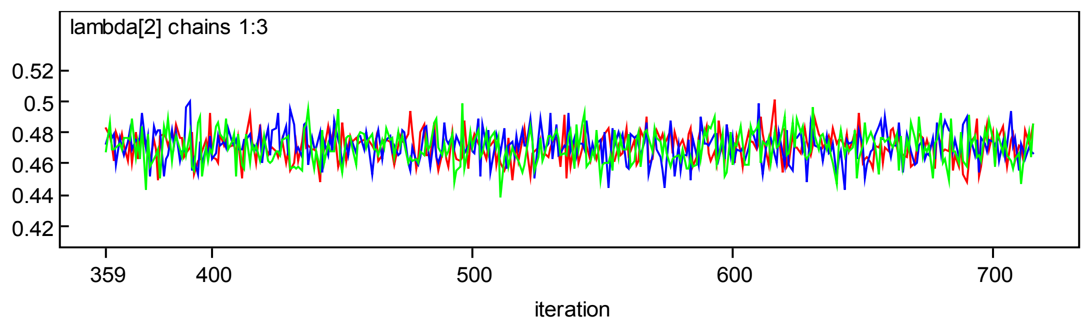

Figure 20. Trace Plot for Kindergarten Lottery Sample Math Lambda 2

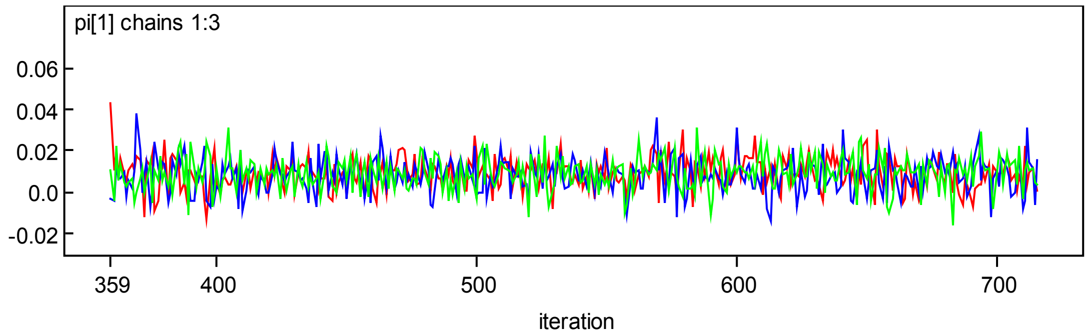

Figure 21. Trace Plot for Kindergarten Lottery Sample Math Pi 1

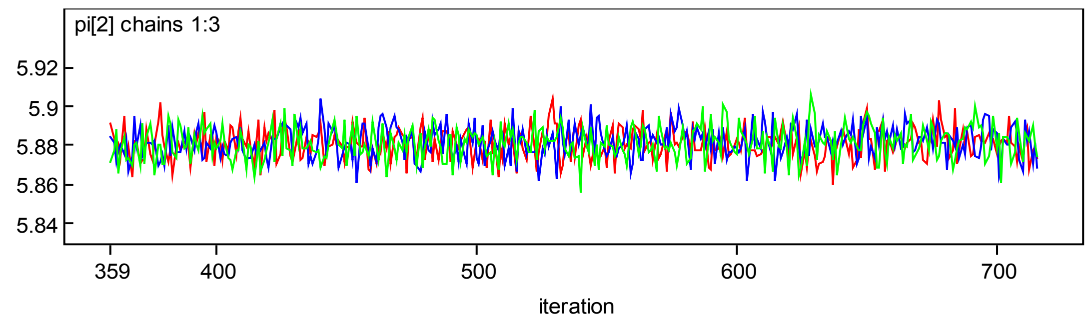

Figure 22. Trace Plot for Kindergarten Lottery Sample Math Pi 2

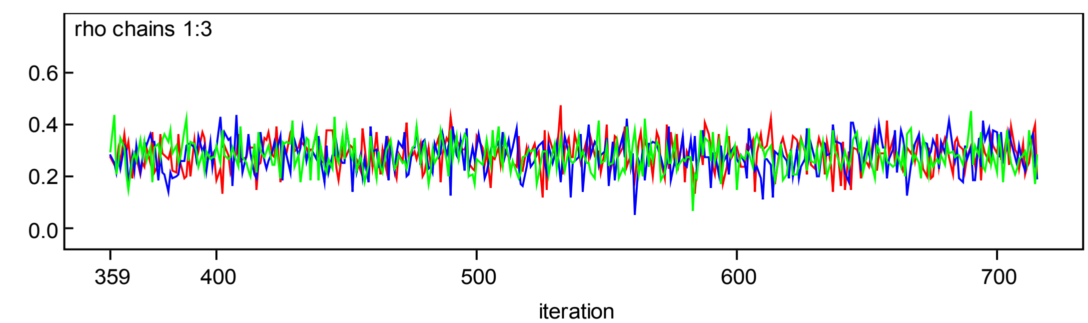

Figure 23. Trace Plot for Kindergarten Lottery Sample Math Rho

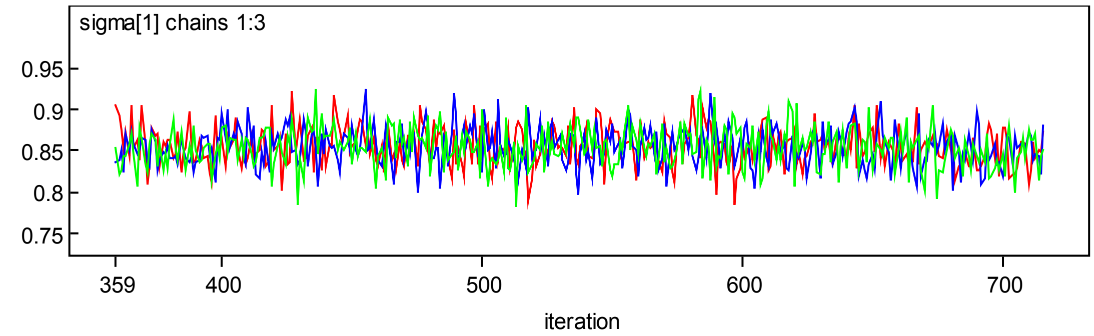

Figure 24. Trace Plot for Kindergarten Lottery Sample Math Sigma 1

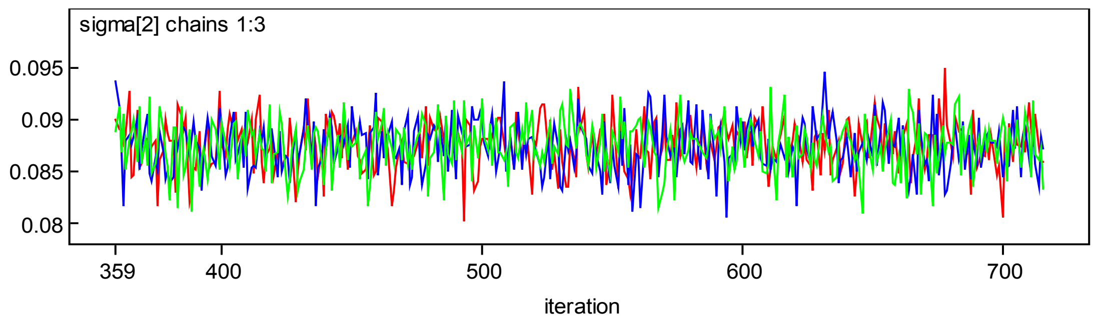

Figure 24. Trace Plot for Kindergarten Lottery Sample Math Sigma 2

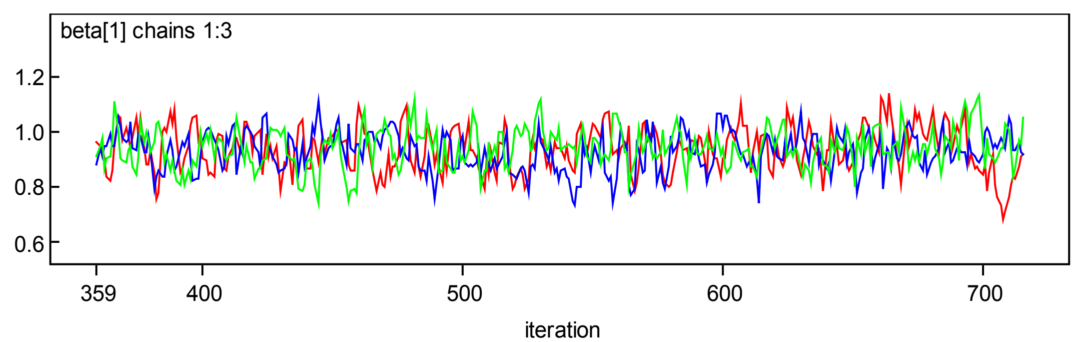

Figure 25. Trace Plot for Kindergarten Lottery Sample Reading Intercept

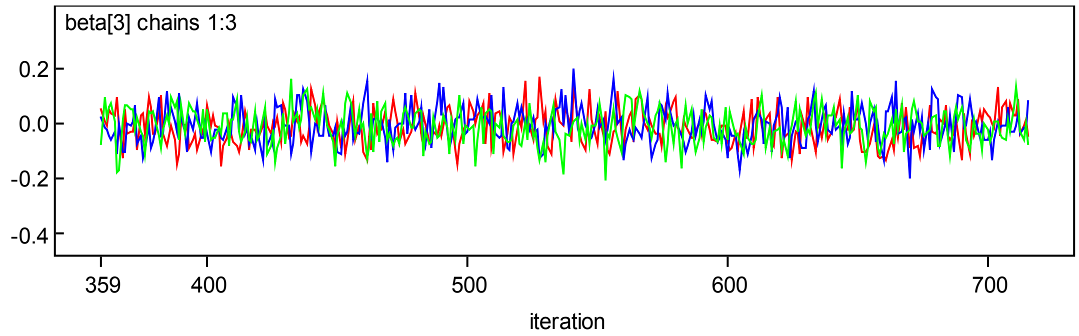

Figure 26. Trace Plot for Kindergarten Lottery Sample Reading Year 07

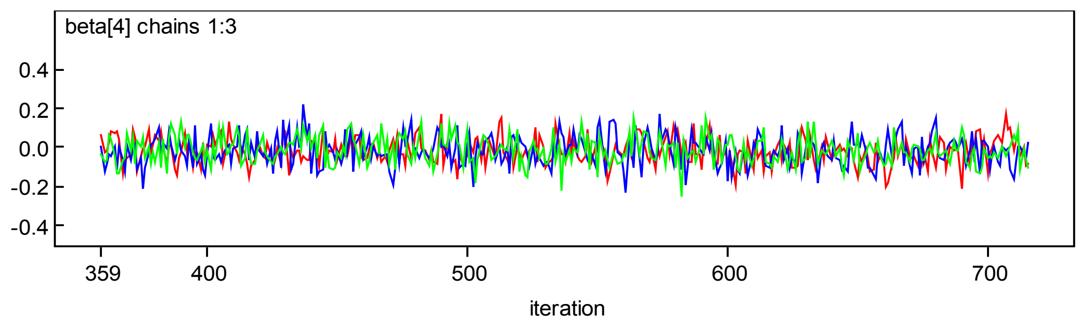

Figure 27. Trace Plot for Kindergarten Lottery Sample Reading Year 08

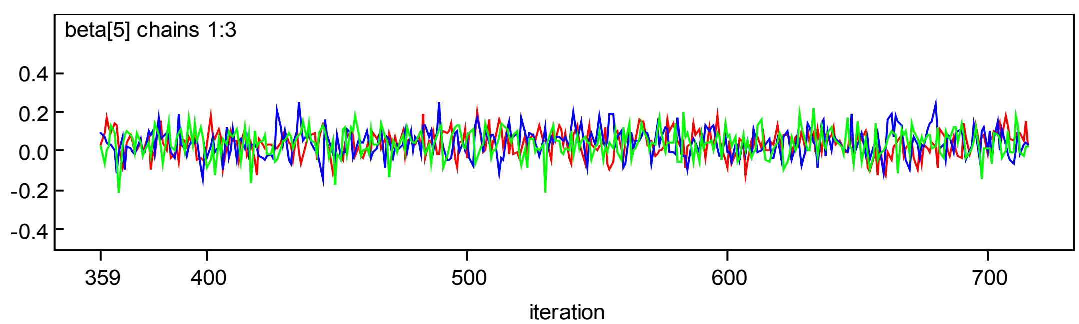

Figure 28. Trace Plot for Kindergarten Lottery Sample Reading Year 09

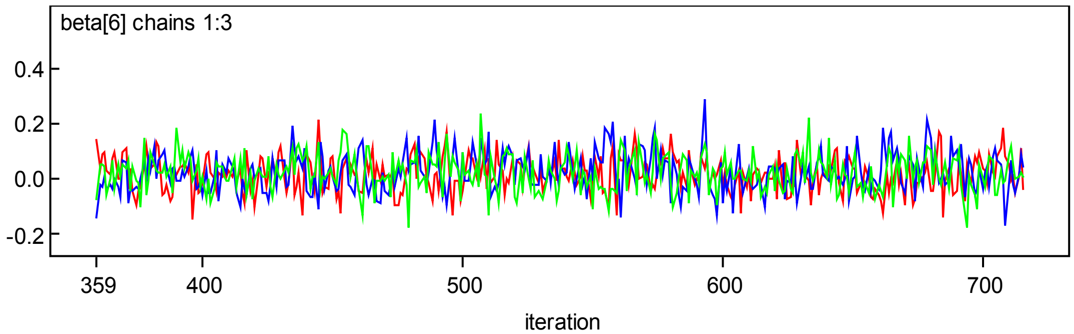

Figure 29. Trace Plot for Kindergarten Lottery Sample Reading Year 10

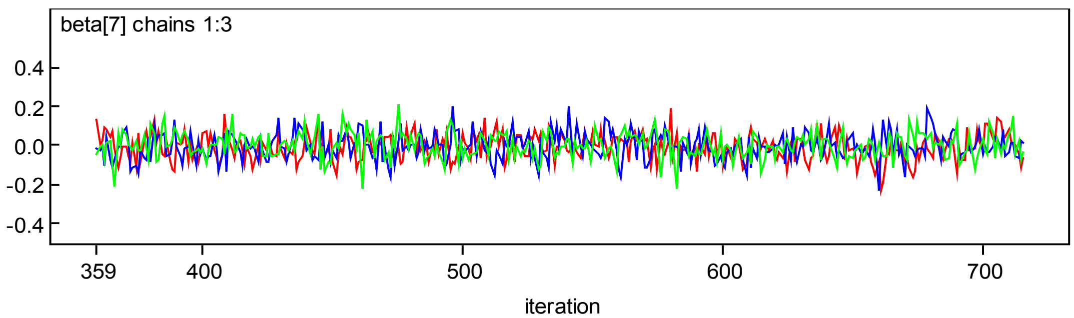

Figure 30. Trace Plot for Kindergarten Lottery Sample Reading Year 11

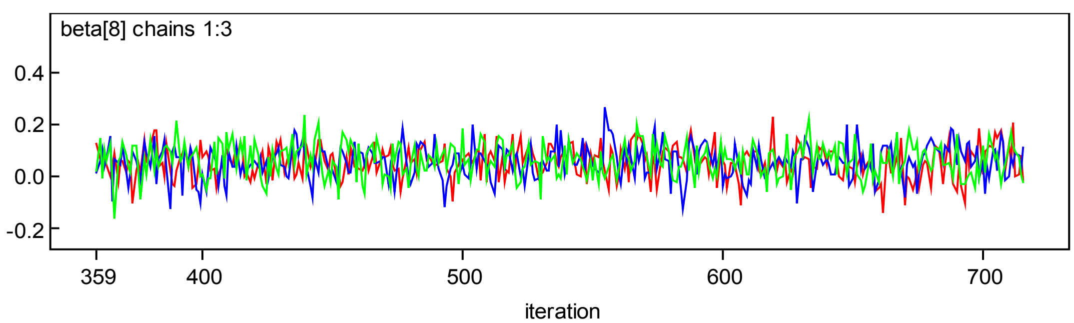

Figure 31. Trace Plot for Kindergarten Lottery Sample Reading Year 12

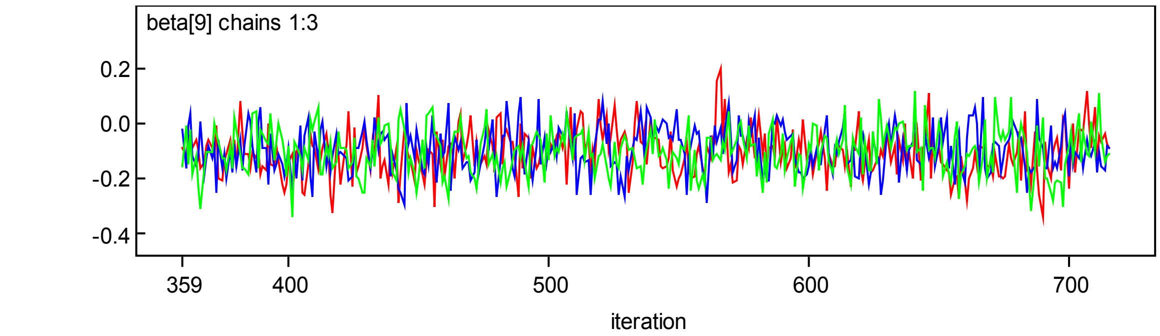

Figure 32. Trace Plot for Kindergarten Lottery Sample Reading Year 13

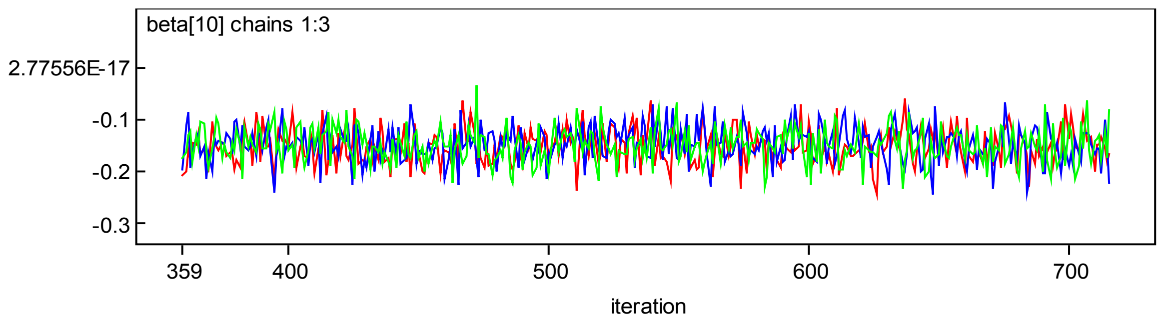

Figure 33. Trace Plot for Kindergarten Lottery Sample Reading Male

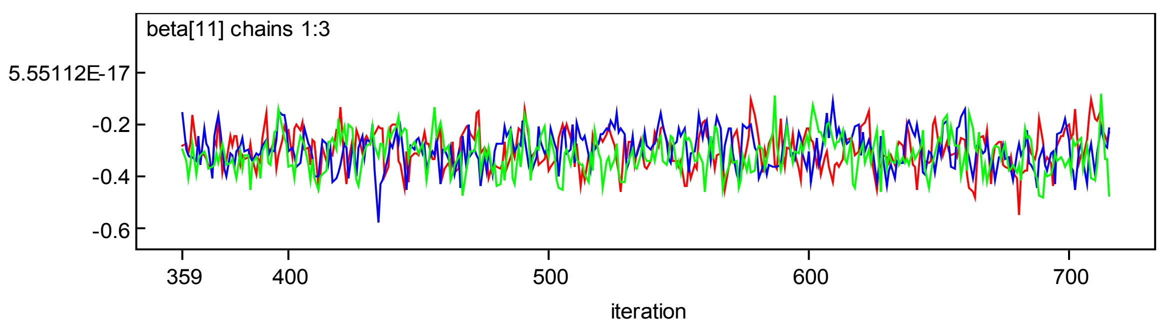

Figure 34. Trace Plot for Kindergarten Lottery Sample Reading Black

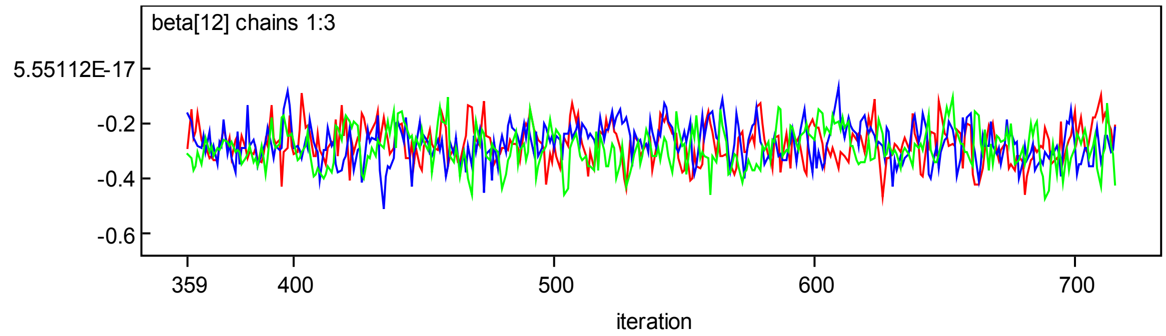

Figure 35. Trace Plot for Kindergarten Lottery Sample Reading Hispanic

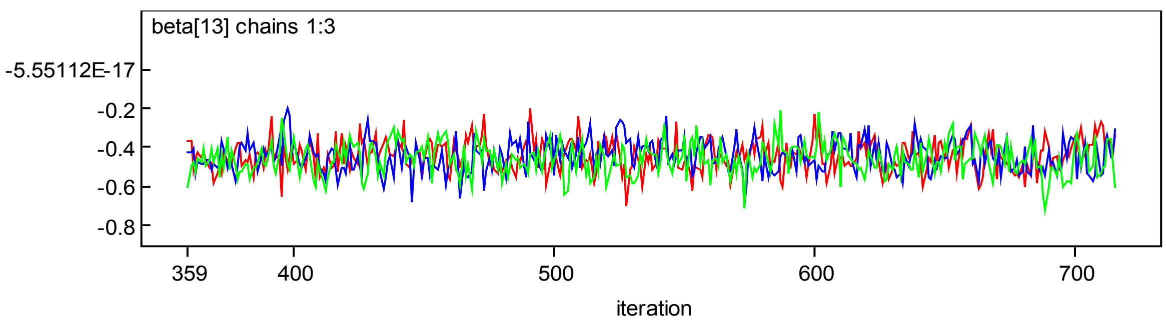

Figure 36. Trace Plot for Kindergarten Lottery Sample Reading Asian

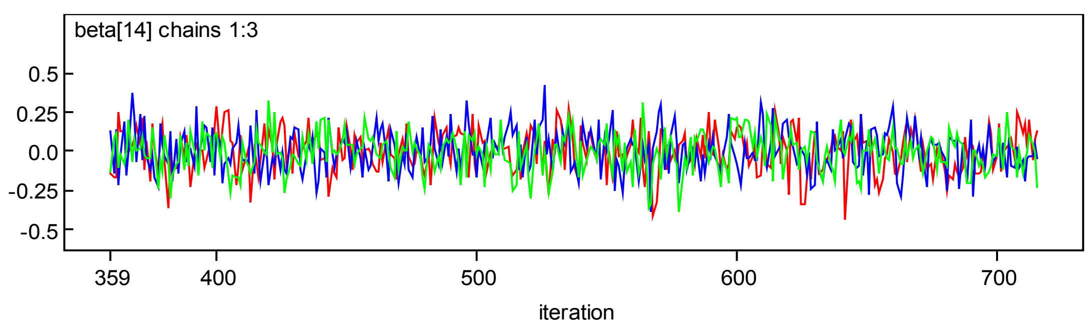

Figure 37. Trace Plot for Kindergarten Lottery Sample Reading Mixed

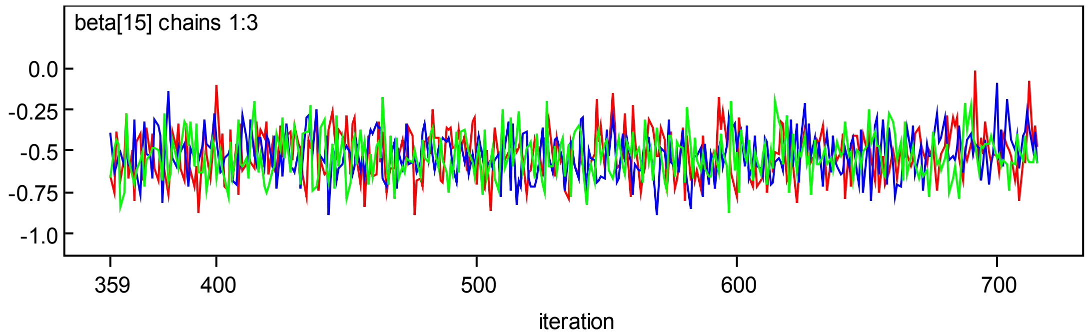

Figure 38. Trace Plot for Kindergarten Lottery Sample Reading Sped 1

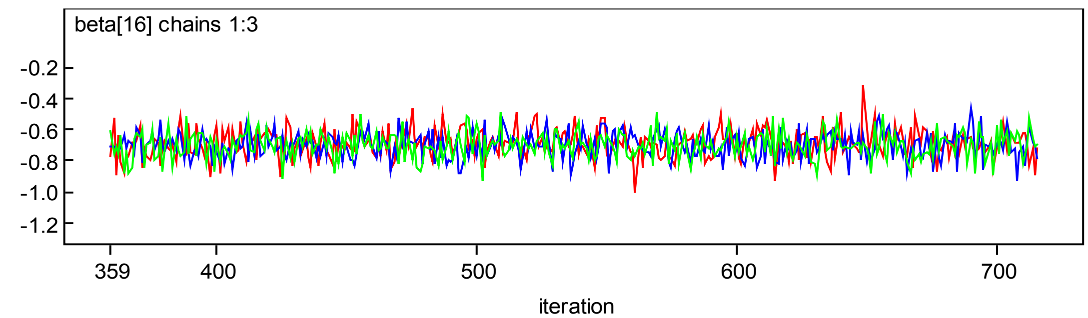

Figure 39. Trace Plot for Kindergarten Lottery Sample Reading Sped 2

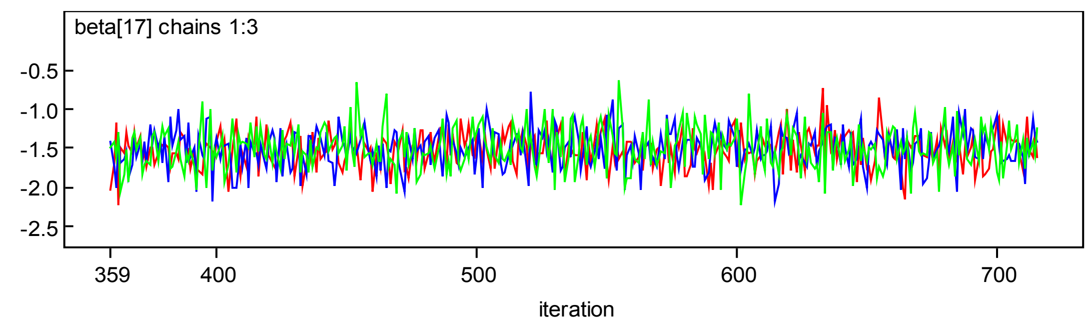

Figure 40. Trace Plot for Kindergarten Lottery Sample Reading Sped 3

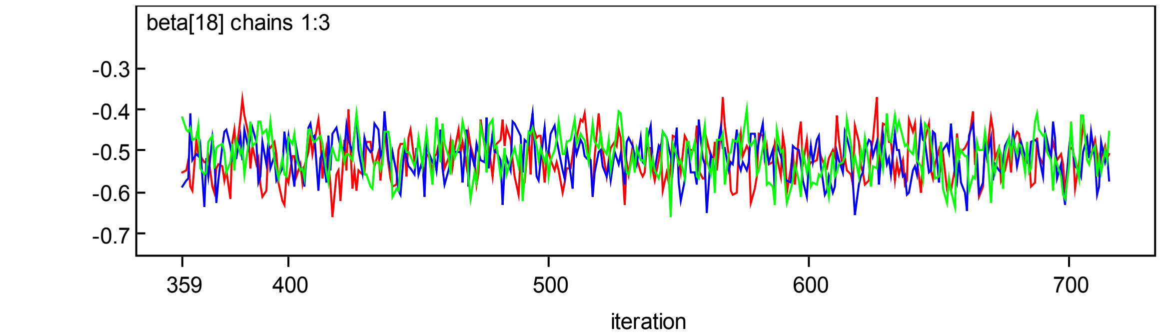

Figure 41. Trace Plot for Kindergarten Lottery Sample Reading Free Lunch

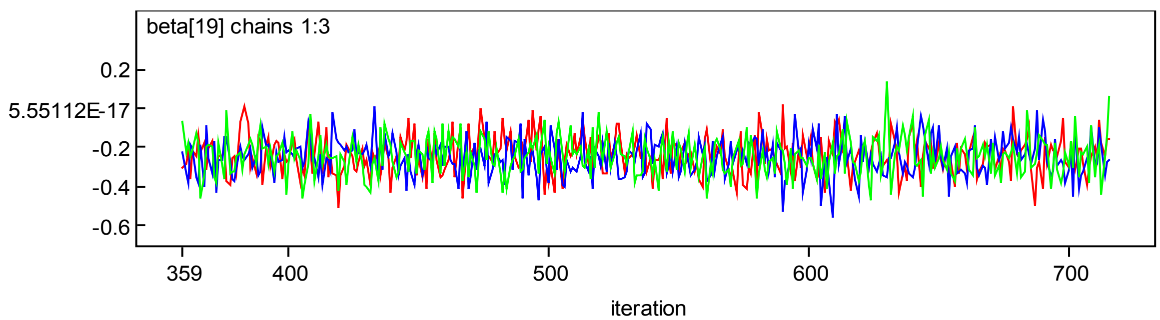

Figure 42. Trace Plot for Kindergarten Lottery Sample Reading Reduced Lunch

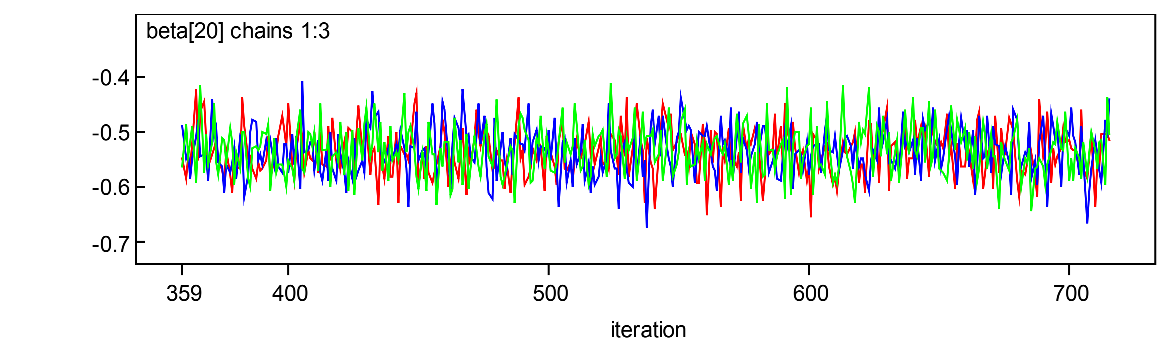

Figure 43. Trace Plot for Kindergarten Lottery Sample Reading Bilingual

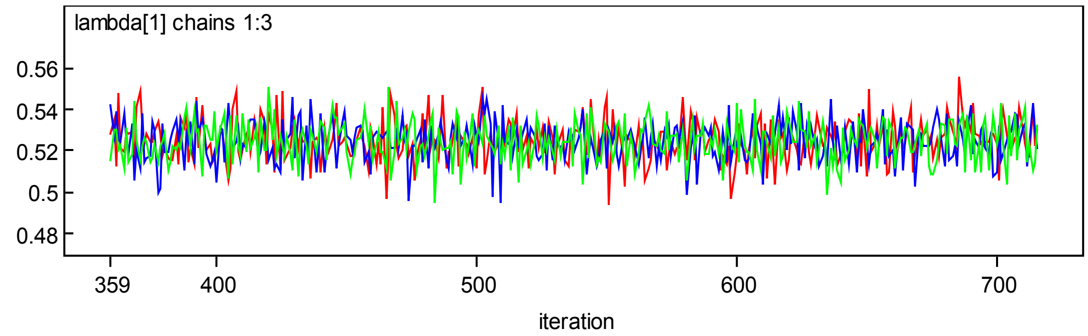

Figure 44. Trace Plot for Kindergarten Lottery Sample Reading Lambda 1

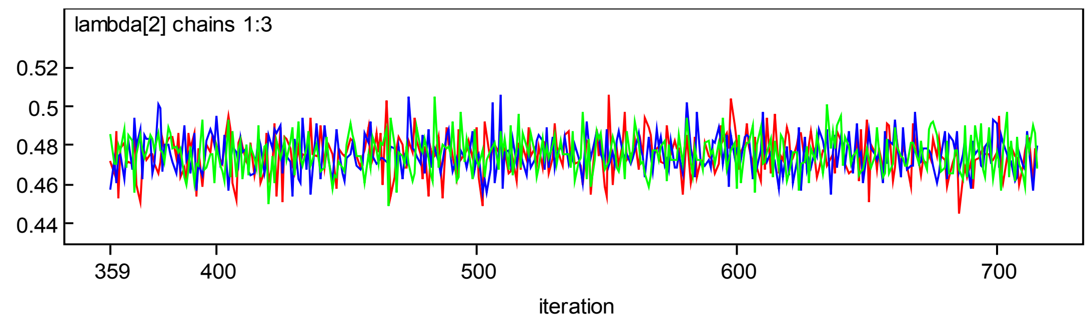

Figure 45. Trace Plot for Kindergarten Lottery Sample Reading Lambda 2

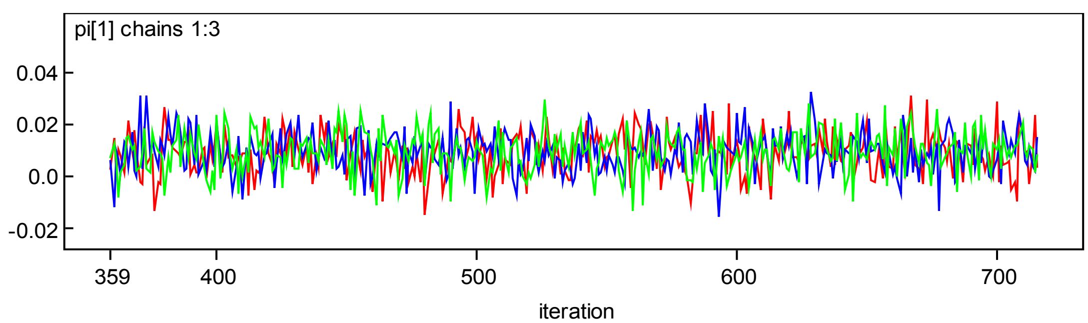

Figure 46. Trace Plot for Kindergarten Lottery Sample Reading Pi 1

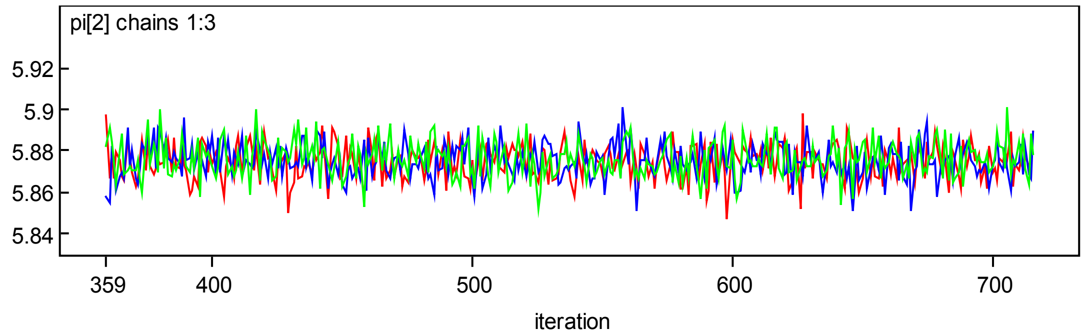

Figure 47. Trace Plot for Kindergarten Lottery Sample Reading Pi 2

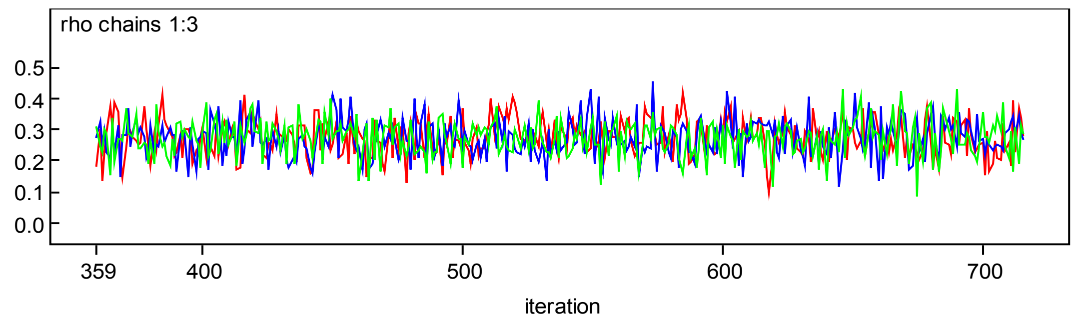

Figure 48. Trace Plot for Kindergarten Lottery Sample Reading Rho

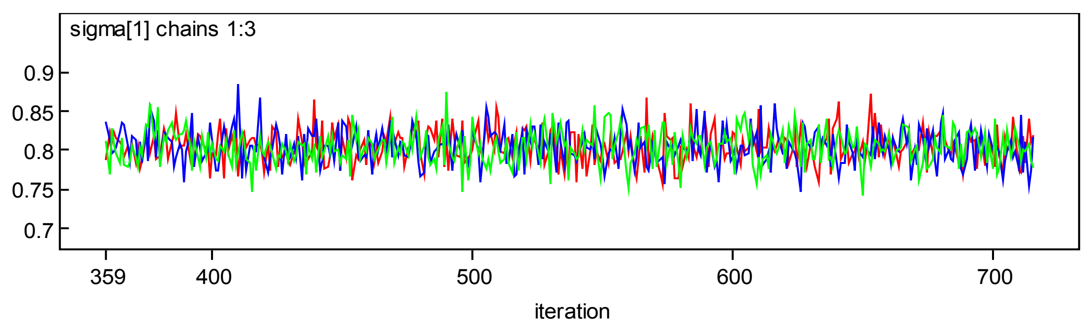

Figure 95. Trace Plot for Kindergarten Lottery Sample Reading Sigma 1

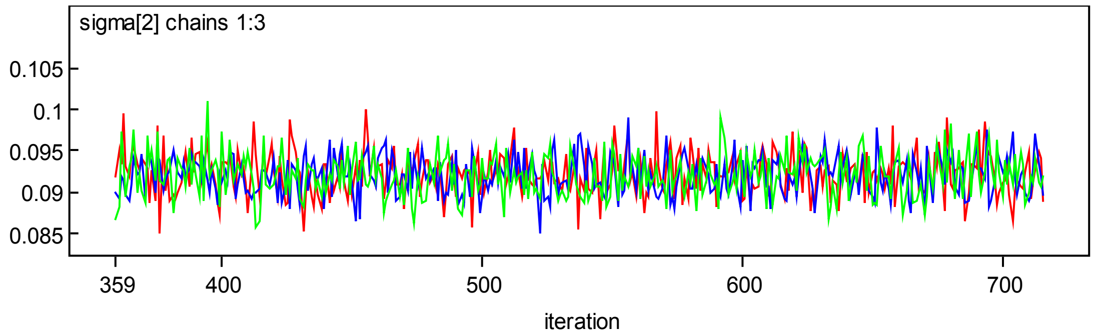

Figure 49. Trace Plot for Kindergarten Lottery Sample Reading Sigma 2

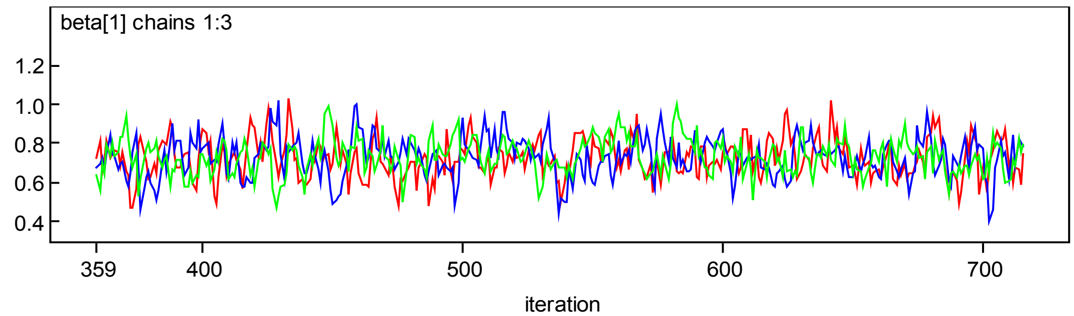

Figure 50. Trace Plot for Grade 3 Lottery Sample Math Intercept

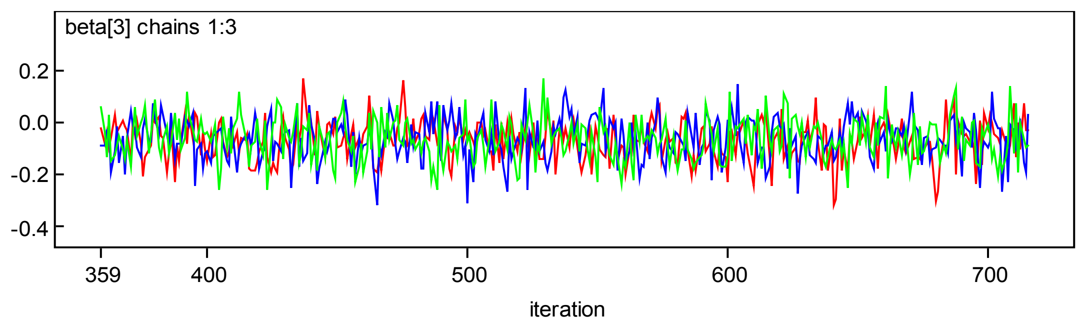

Figure 51. Trace Plot for Grade 3 Lottery Sample Math Year 10

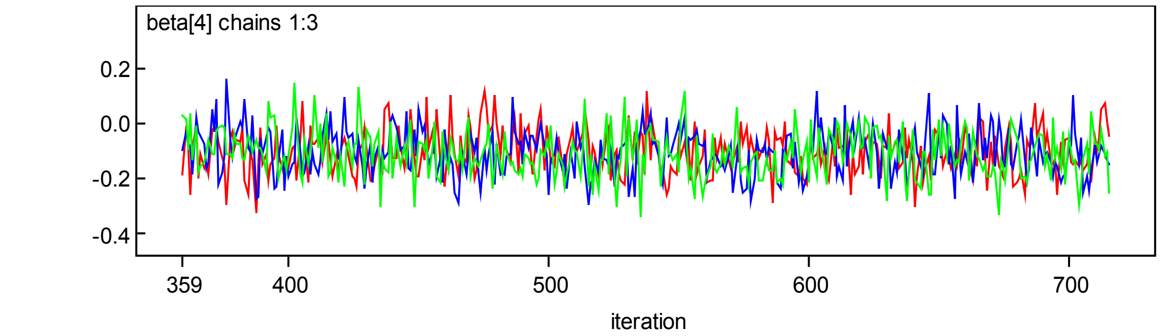

Figure 52. Trace Plot for Grade 3 Lottery Sample Math Year 11

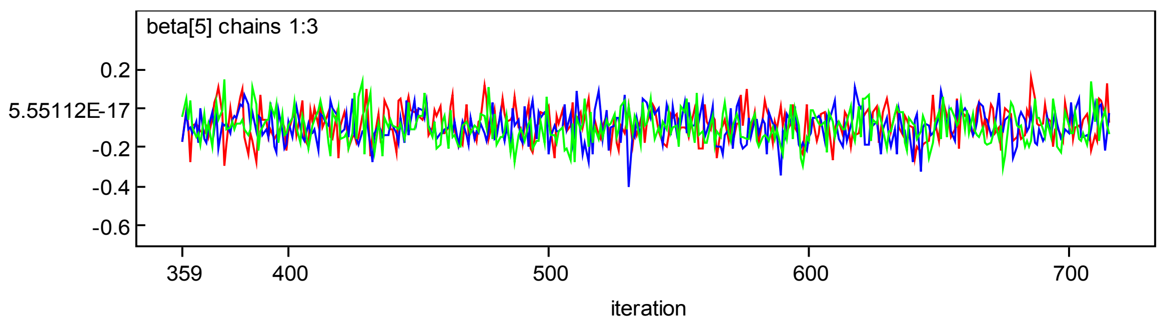

Figure 53. Trace Plot for Grade 3 Lottery Sample Math Year 12

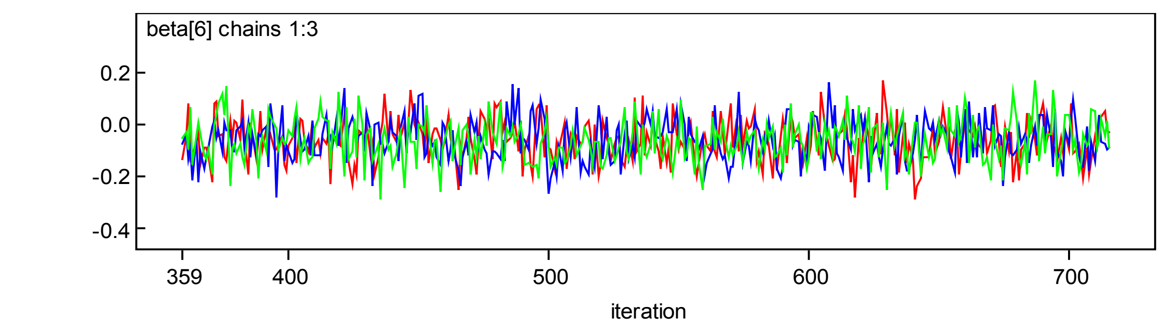

Figure 54. Trace Plot for Grade 3 Lottery Sample Math Year 13

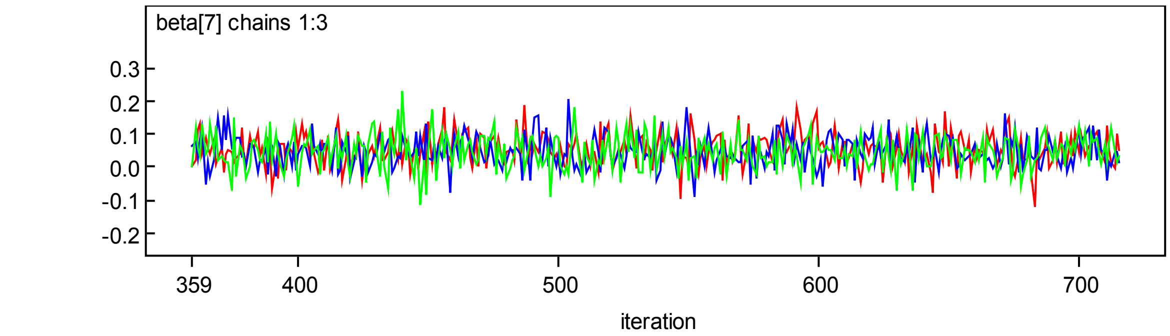

Figure 55. Trace Plot for Grade 3 Lottery Sample Math Male

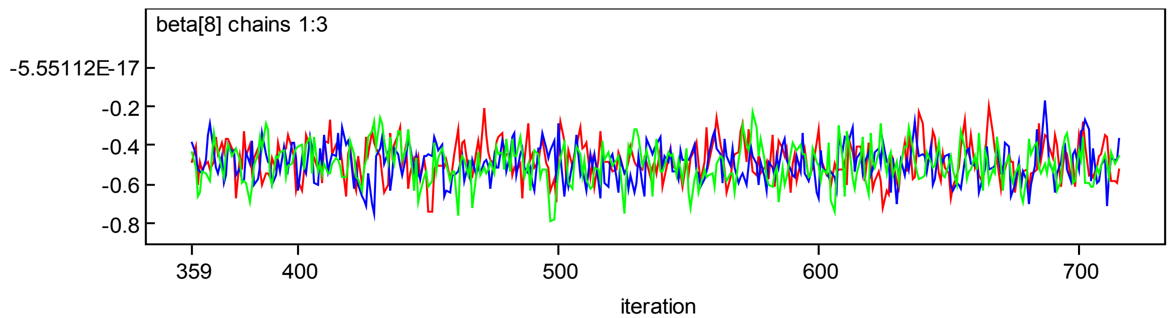

Figure 56. Trace Plot for Grade 3 Lottery Sample Math Black

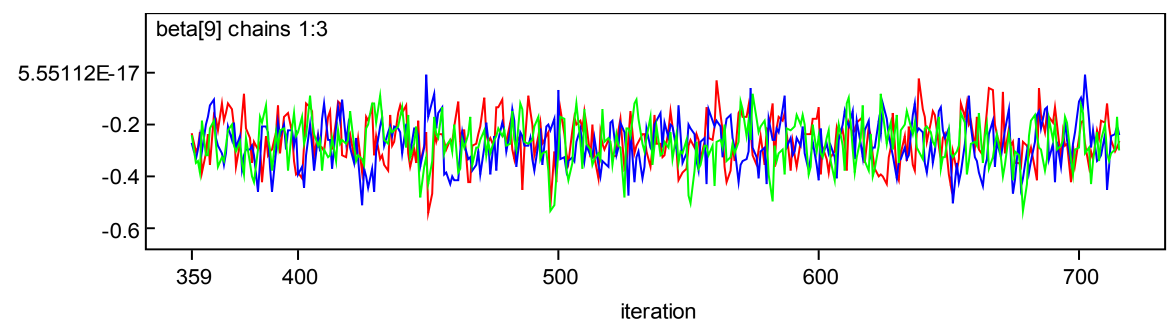

Figure 57. Trace Plot for Grade 3 Lottery Sample Math Hispanic

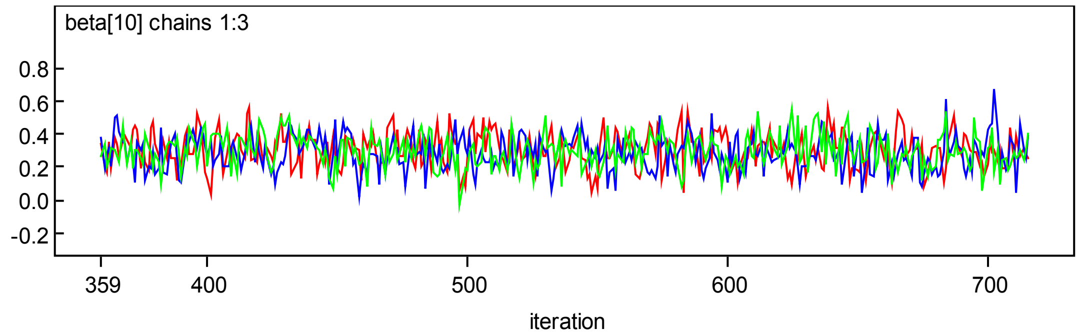

Figure 58. Trace Plot for Grade 3 Lottery Sample Math Asian

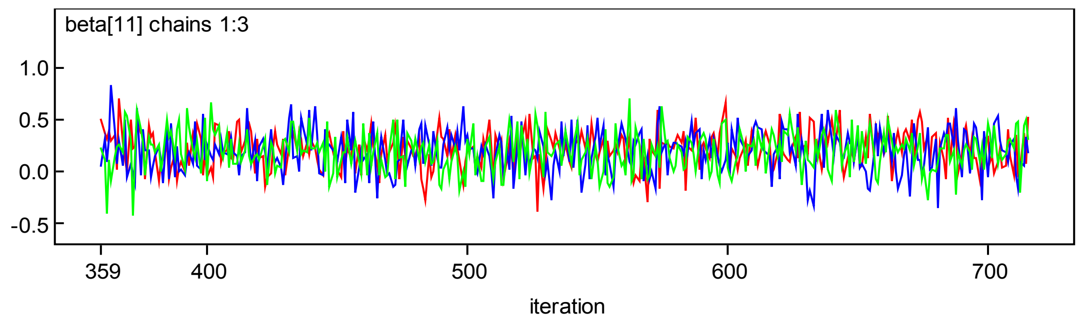

Figure 59. Trace Plot for Grade 3 Lottery Sample Math Mixed

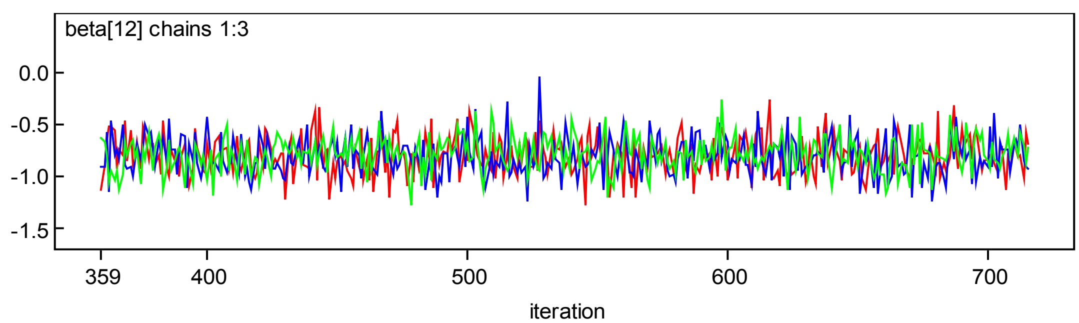

Figure 60. Trace Plot for Grade 3 Lottery Sample Math Sped 1

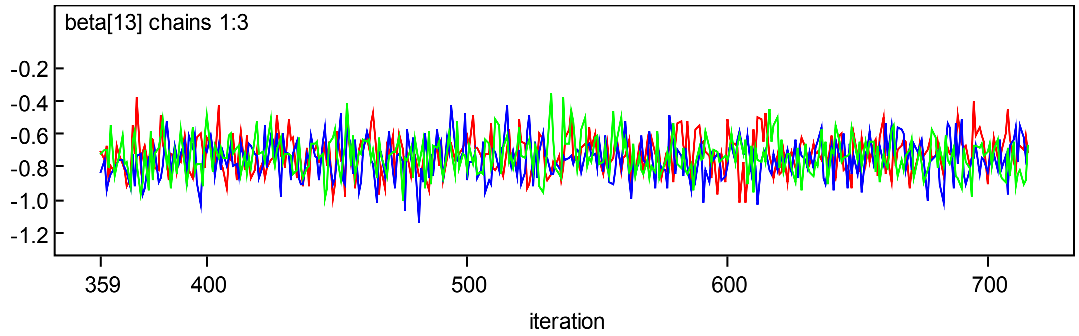

Figure 61. Trace Plot for Grade 3 Lottery Sample Math Sped 2

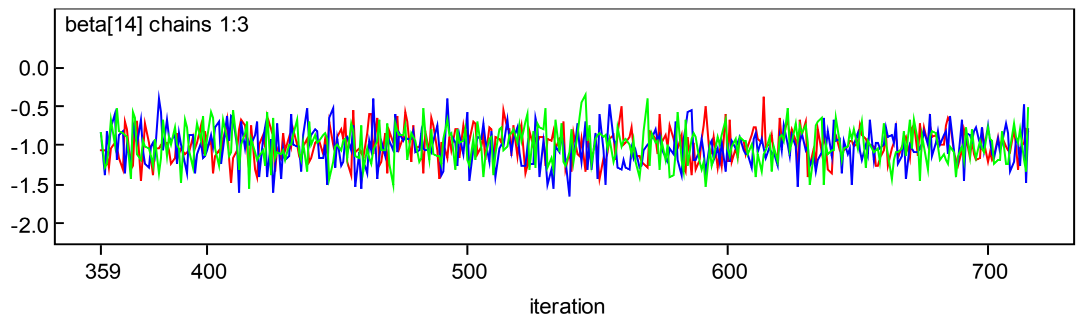

Figure 62. Trace Plot for Grade 3 Lottery Sample Math Sped 3

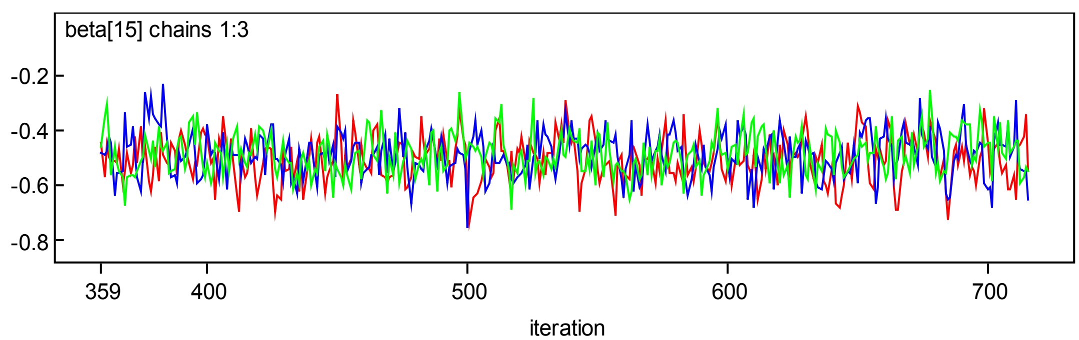

Figure 63. Trace Plot for Grade 3 Lottery Sample Math Free Lunch

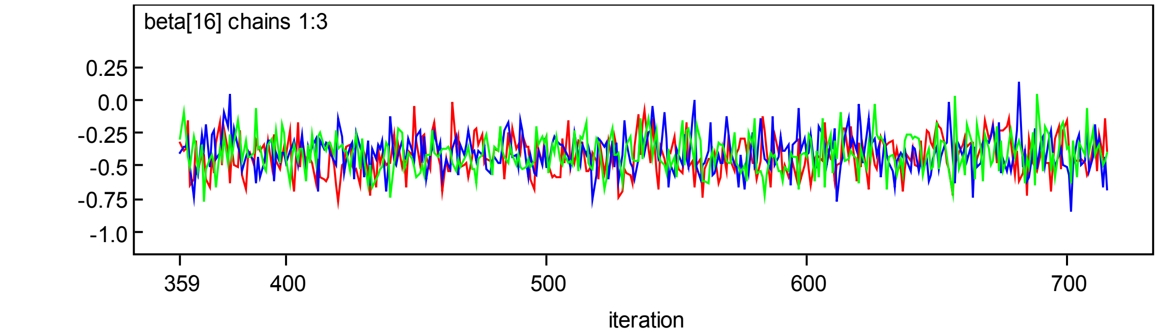

Figure 64. Trace Plot for Grade 3 Lottery Sample Math Reduced Lunch

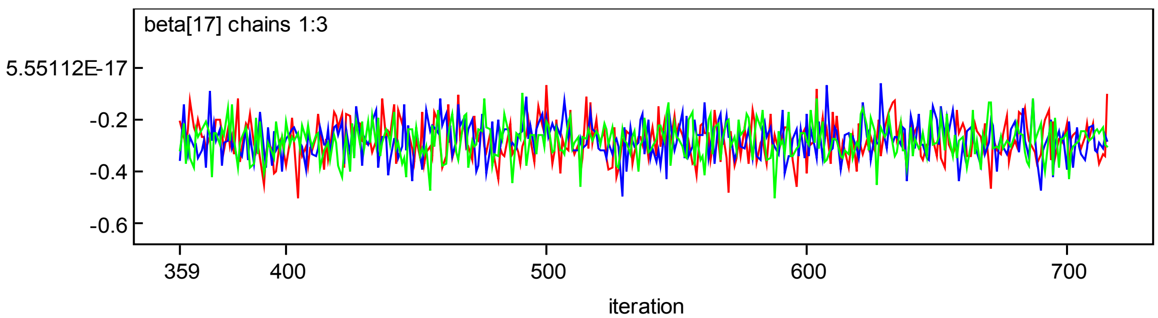

Figure 65. Trace Plot for Grade 3 Lottery Sample Math Bilingual

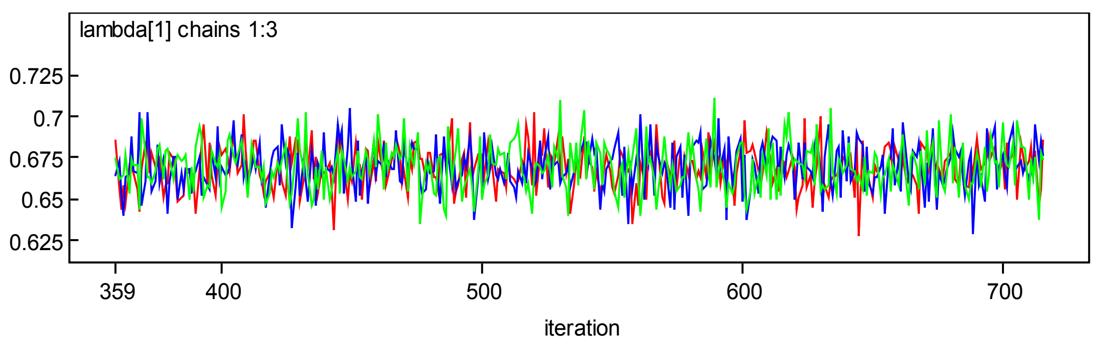

Figure 66. Trace Plot for Grade 3 Lottery Sample Math Lambda 1

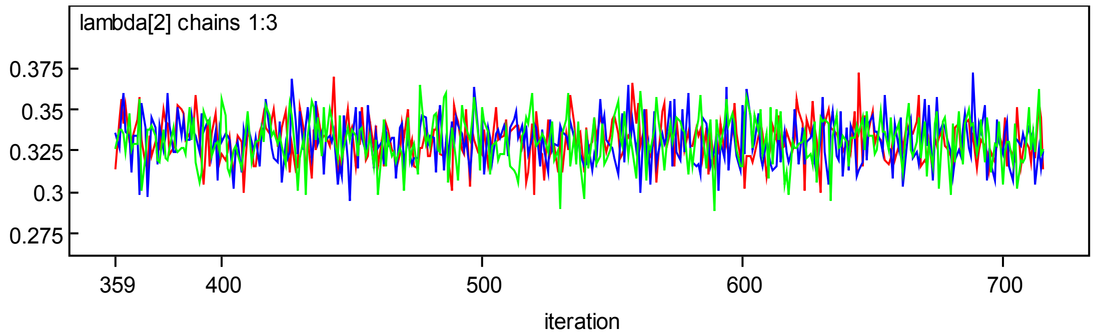

Figure 67. Trace Plot for Grade 3 Lottery Sample Math Lambda 2

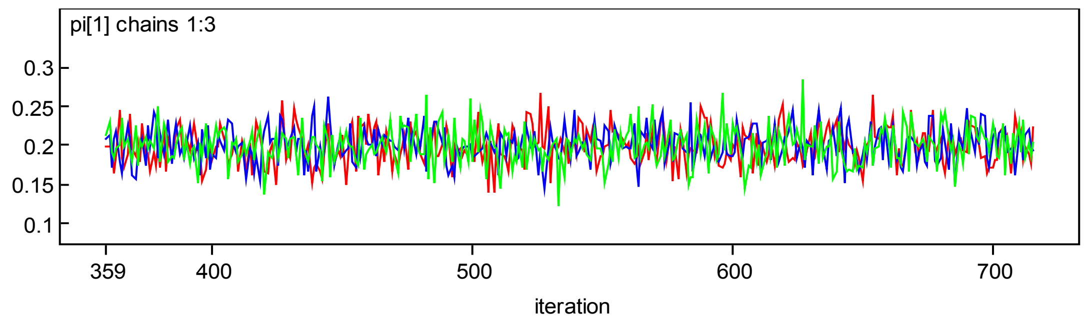

Figure 68. Trace Plot for Grade 3 Lottery Sample Math Pi 1

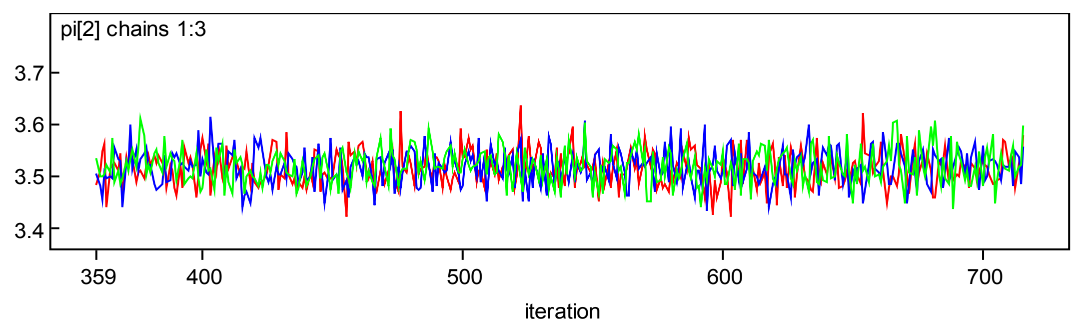

Figure 69. Trace Plot for Grade 3 Lottery Sample Math Pi 2

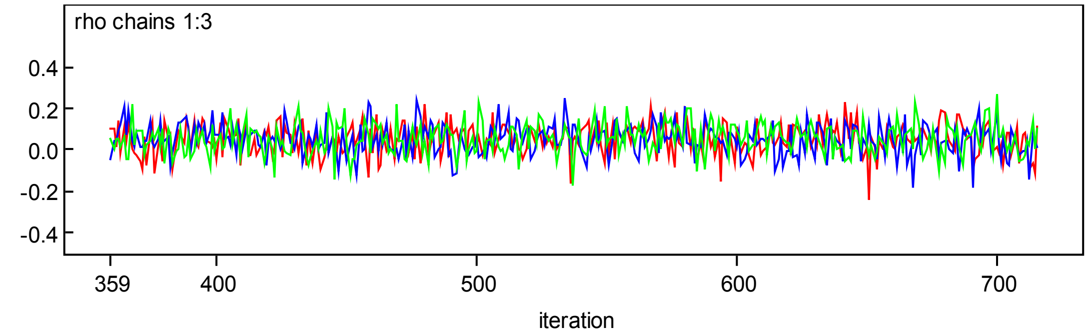

Figure 70. Trace Plot for Grade 3 Lottery Sample Math Rho

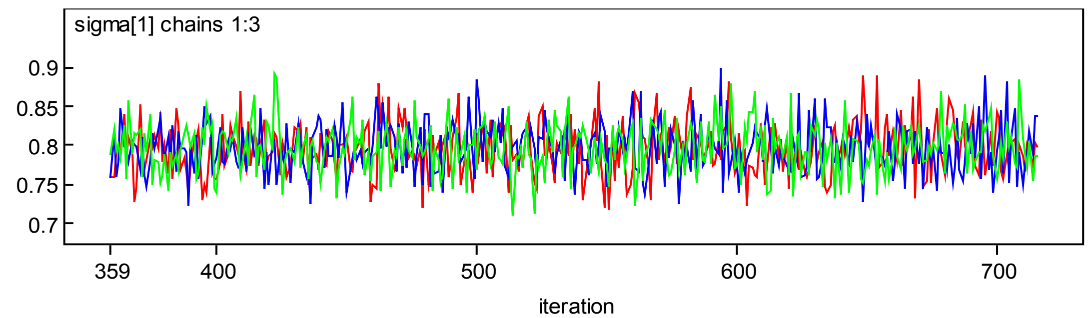

Figure 71. Trace Plot for Grade 3 Lottery Sample Math Sigma 1

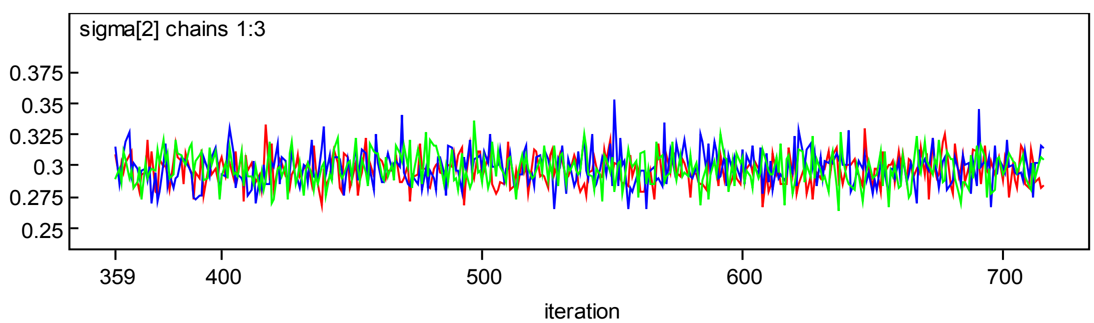

Figure 72. Trace Plot for Grade 3 Lottery Sample Math Sigma 2

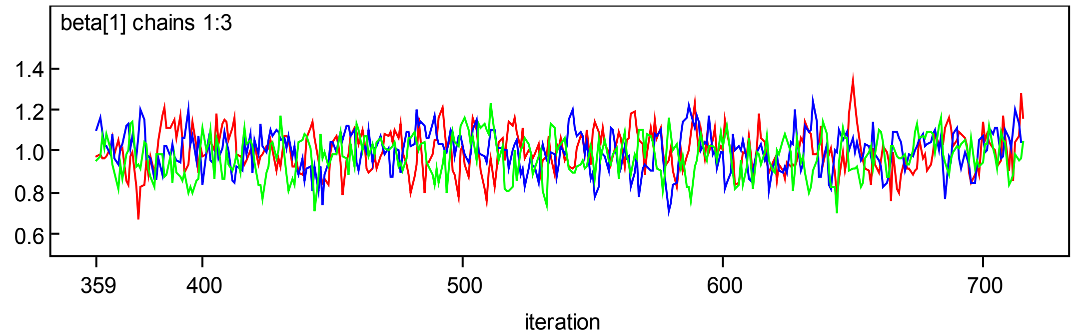

Figure 73. Trace Plot for Grade 3 Lottery Sample ELA Intercept

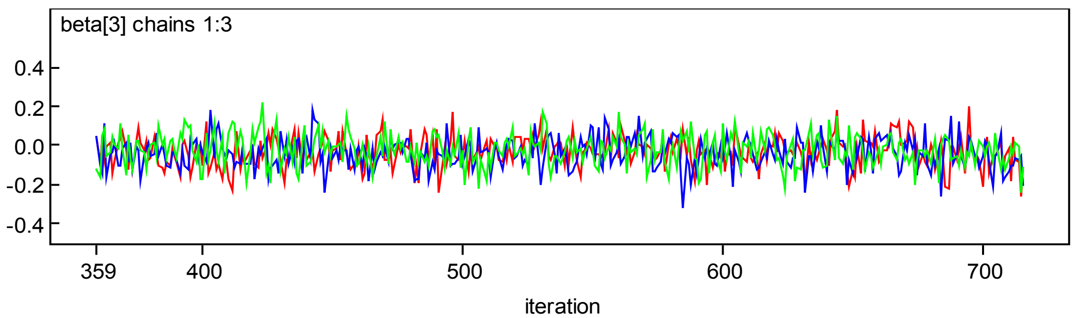

Figure 74. Trace Plot for Grade 3 Lottery Sample ELA Year 10

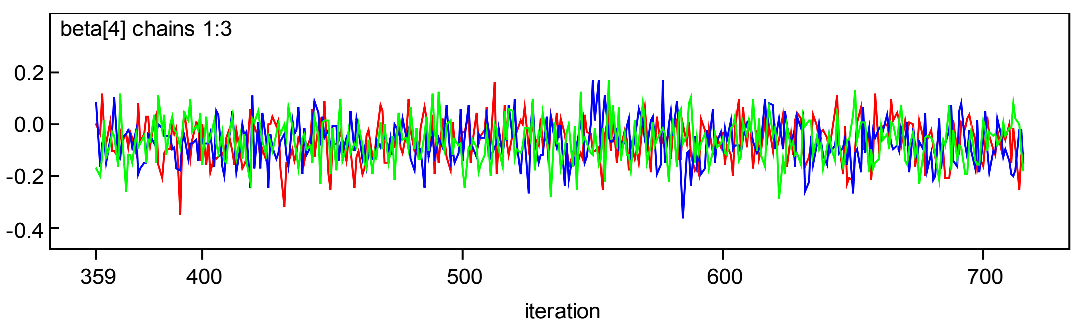

Figure 75. Trace Plot for Grade 3 Lottery Sample ELA Year 11

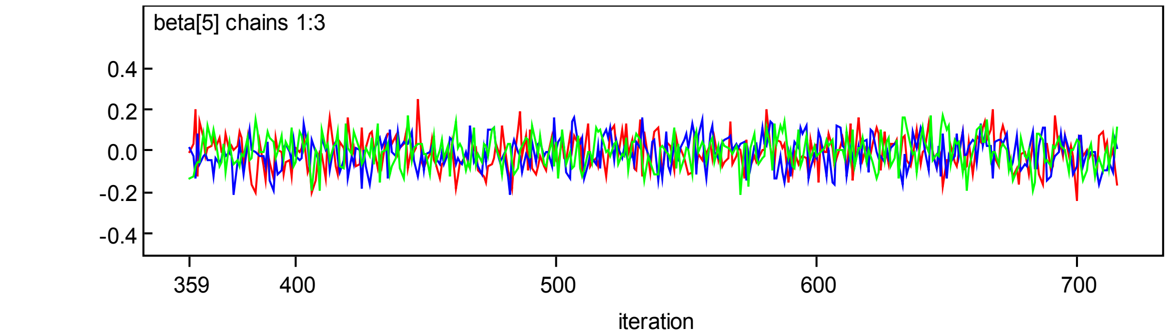

Figure 76. Trace Plot for Grade 3 Lottery Sample ELA Year 12

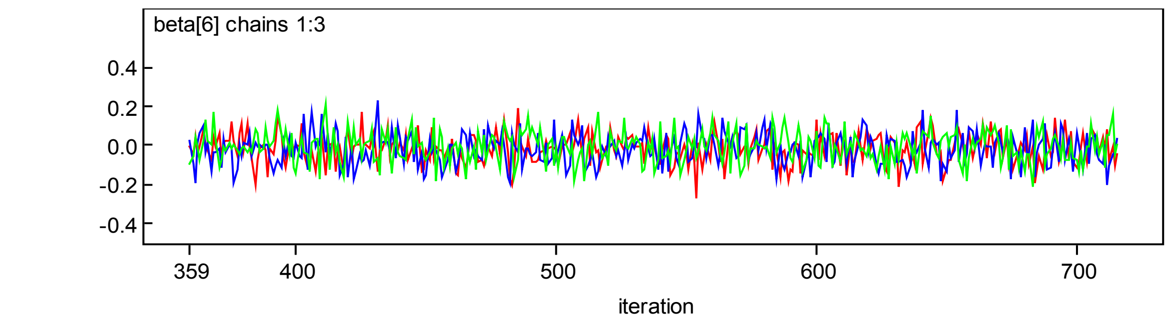

Figure 77. Trace Plot for Grade 3 Lottery Sample ELA Year 13

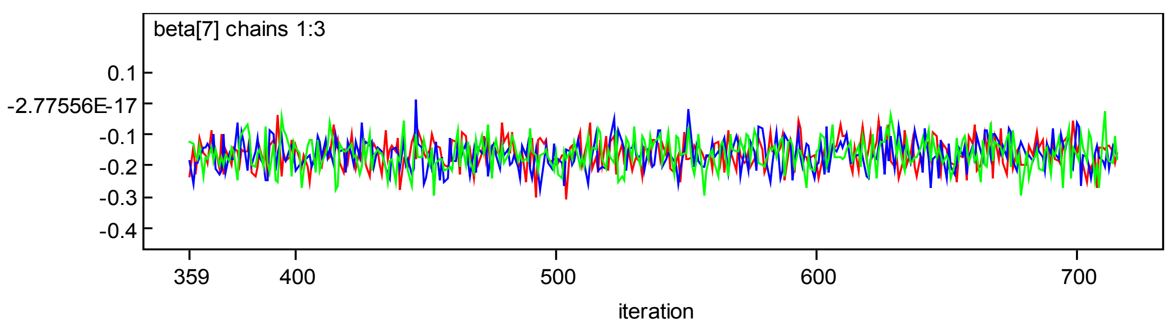

Figure 78. Trace Plot for Grade 3 Lottery Sample ELA Male

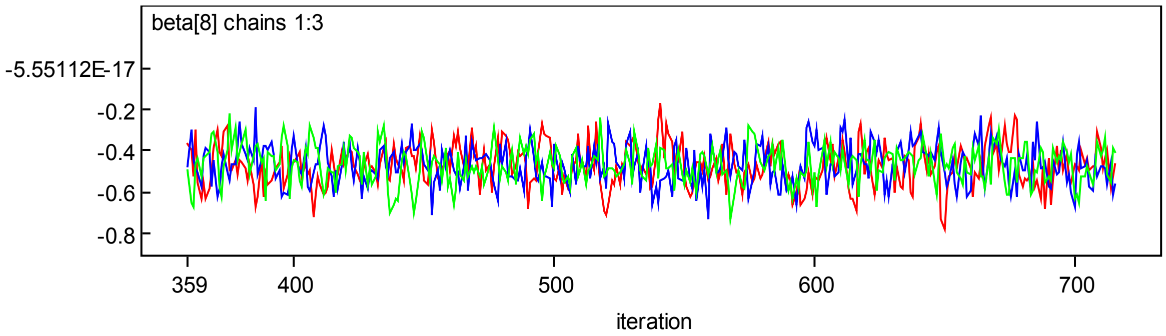

Figure 79. Trace Plot for Grade 3 Lottery Sample ELA Black

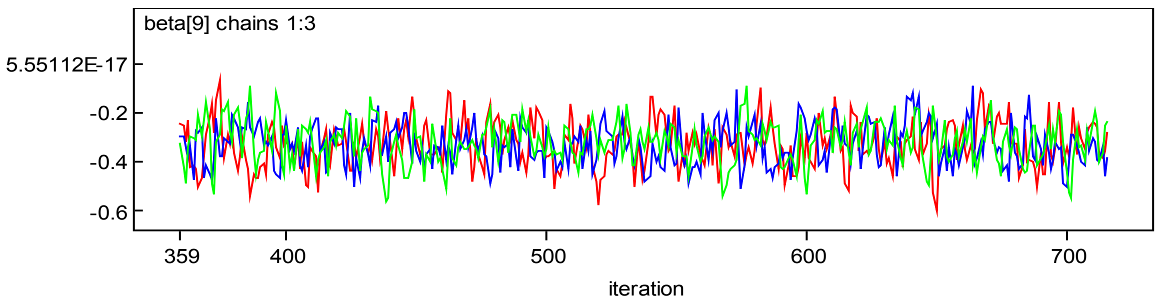

Figure 80. Trace Plot for Grade 3 Lottery Sample ELA Hispanic

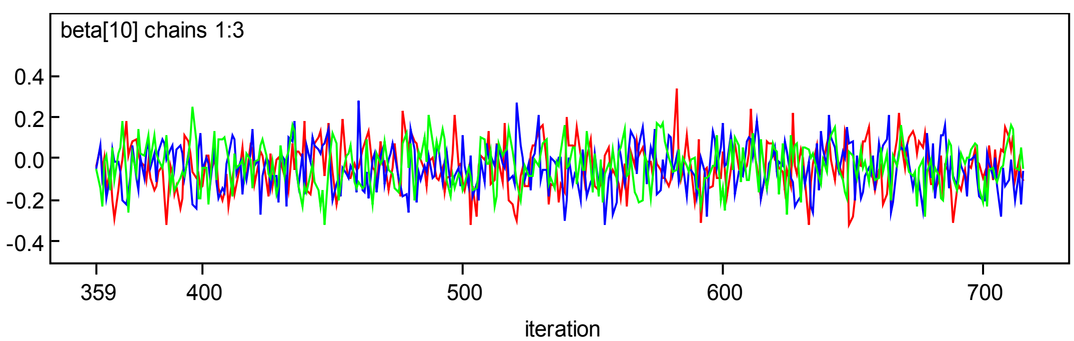

Figure 81. Trace Plot for Grade 3 Lottery Sample ELA Asian

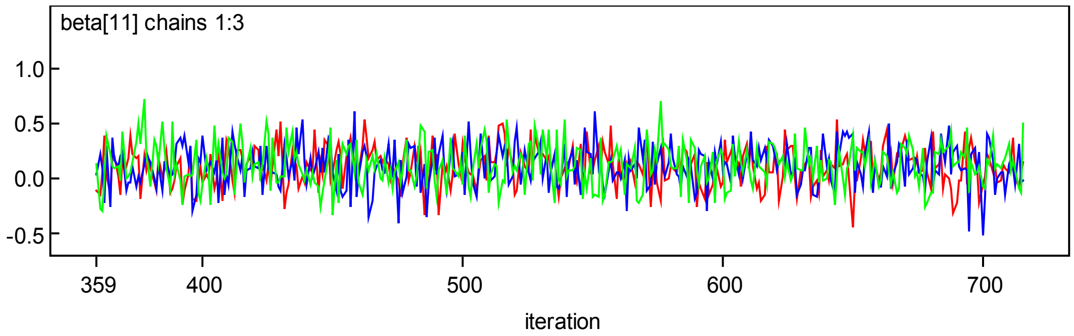

Figure 82. Trace Plot for Grade 3 Lottery Sample ELA Mixed

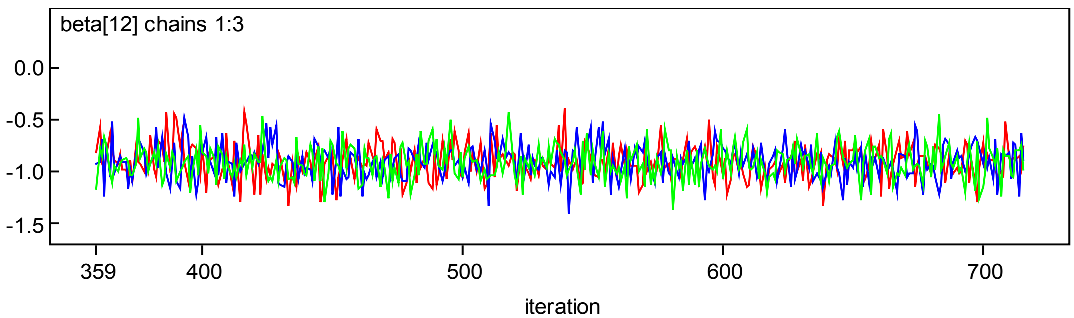

Figure 83. Trace Plot for Grade 3 Lottery Sample ELA Sped 1

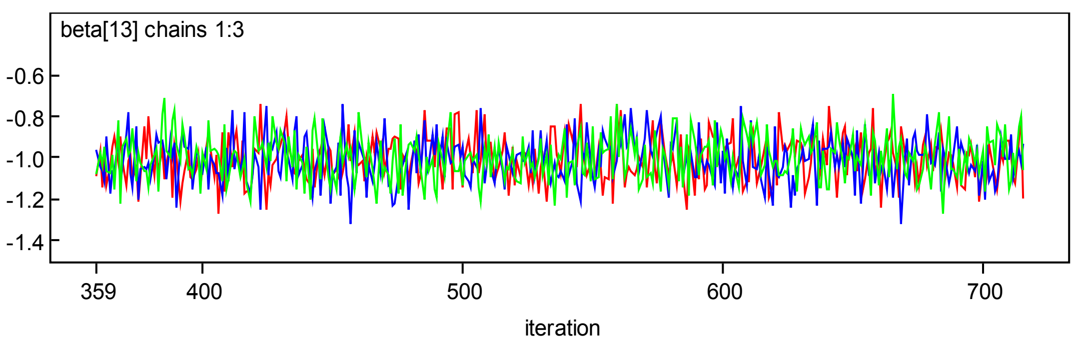

Figure 84. Trace Plot for Grade 3 Lottery Sample ELA Sped 2

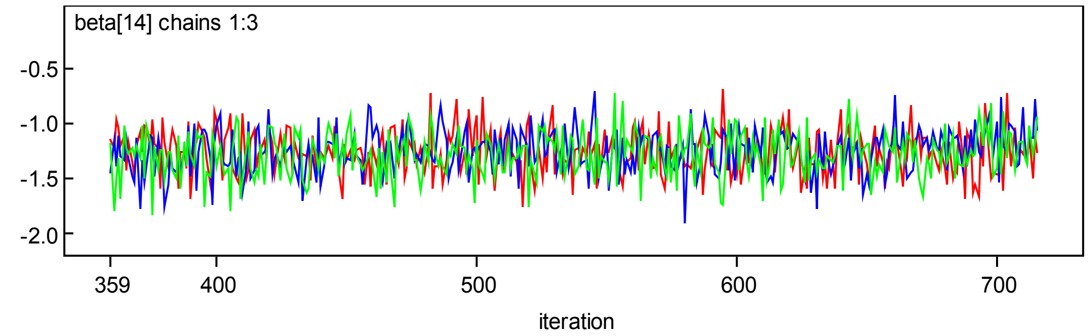

Figure 85. Trace Plot for Grade 3 Lottery Sample ELA Sped 3

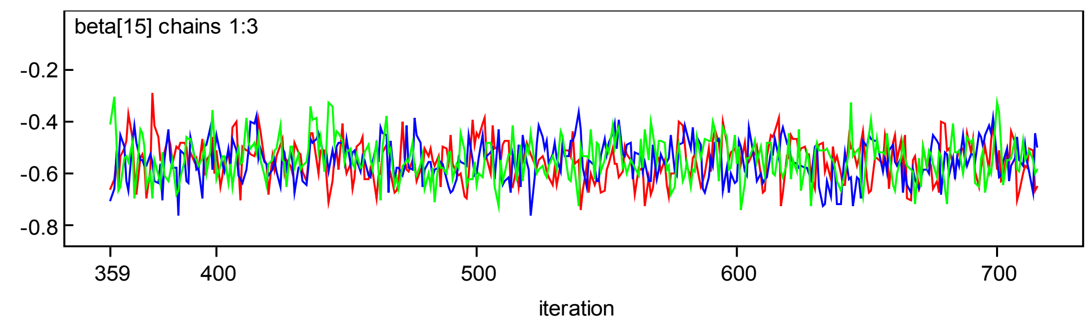

Figure 86. Trace Plot for Grade 3 Lottery Sample ELA Free Lunch

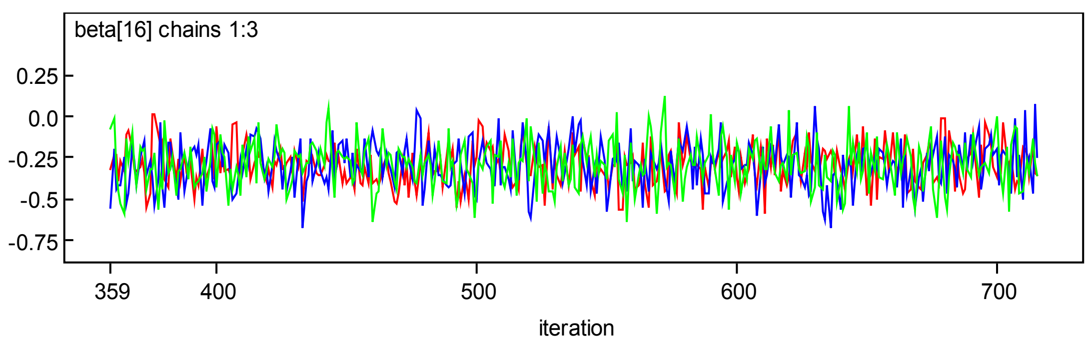

Figure 87. Trace Plot for Grade 3 Lottery Sample ELA Reduced Lunch

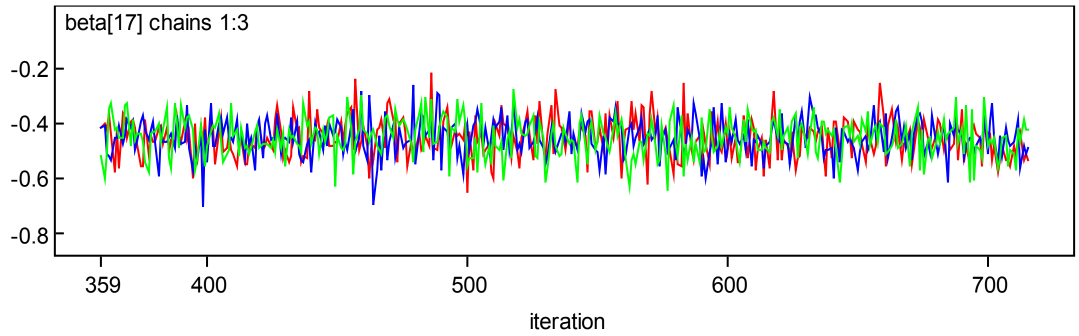

Figure 88. Trace Plot for Grade 3 Lottery Sample ELA Bilingual

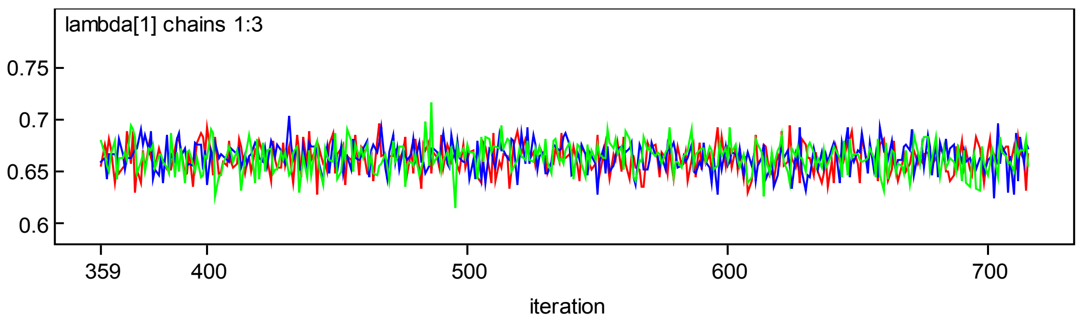

Figure 89. Trace Plot for Grade 3 Lottery Sample ELA Lambda 1

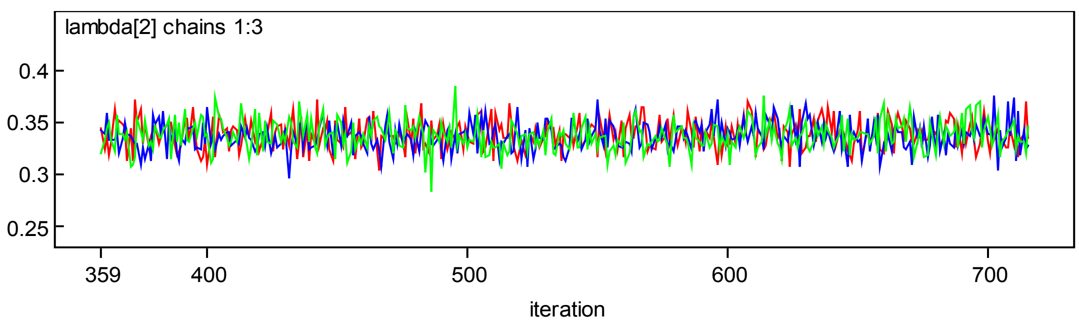

Figure 90. Trace Plot for Grade 3 Lottery Sample ELA Lambda 2

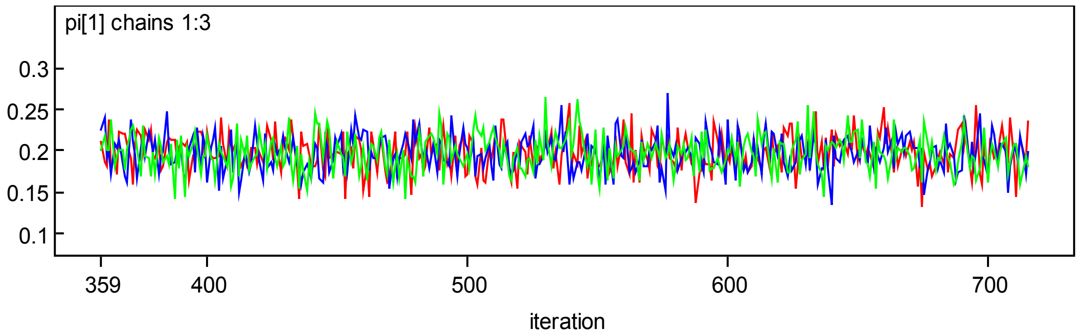

Figure 91. Trace Plot for Grade 3 Lottery Sample ELA Pi 1

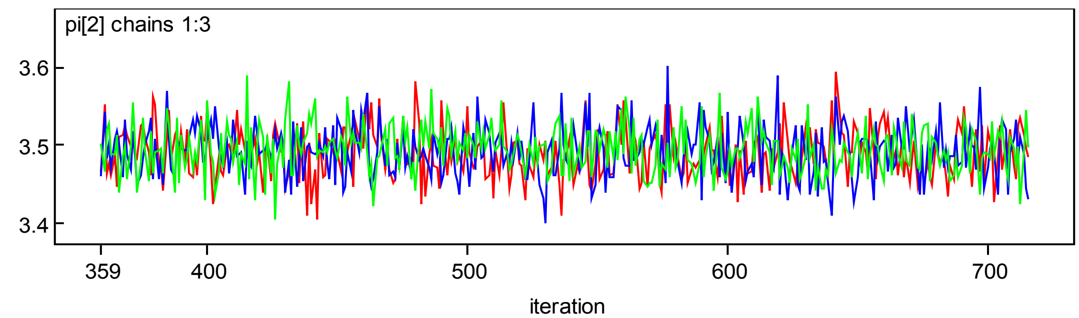

Figure 92. Trace Plot for Grade 3 Lottery Sample ELA Pi 2

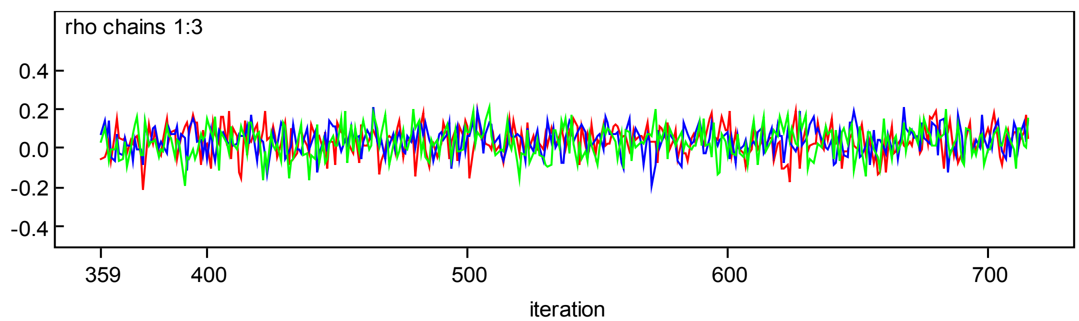

Figure 93. Trace Plot for Grade 3 Lottery Sample ELA Rho

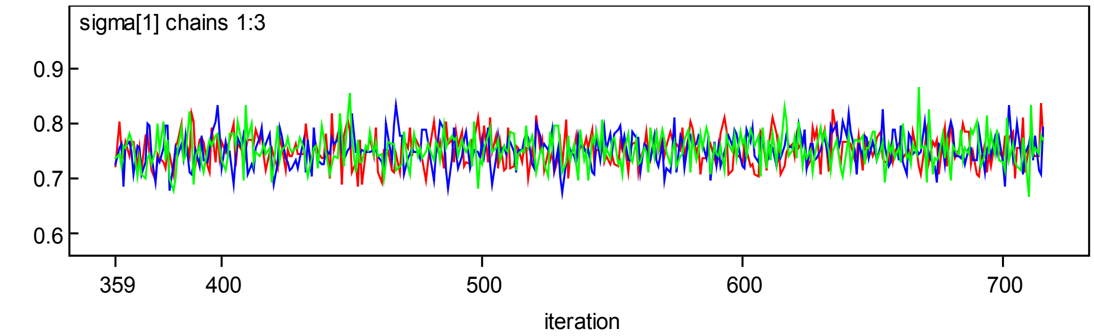

Figure 94. Trace Plot for Grade 3 Lottery Sample ELA Sigma 1

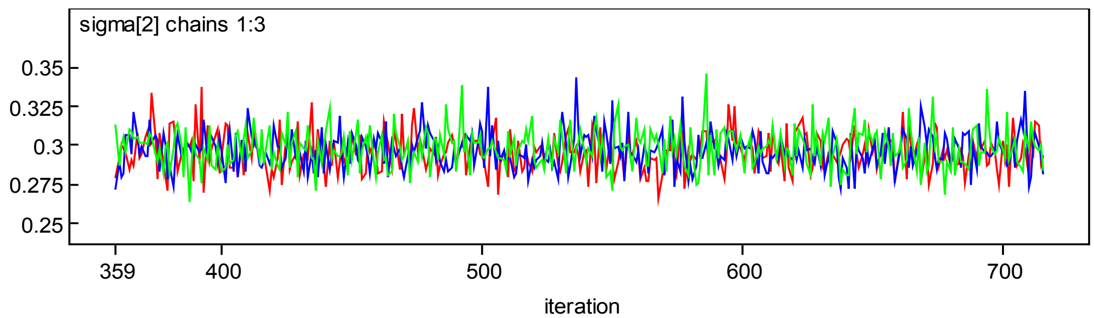

Figure 95. Trace Plot for Grade 3 Lottery Sample ELA Sigma 2

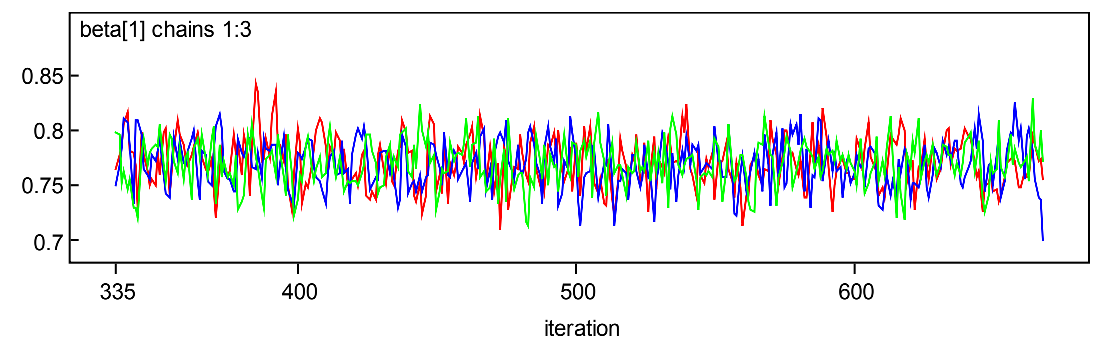

Figure 96. Trace Plot for Kindergarten Full Sample Math Intercept

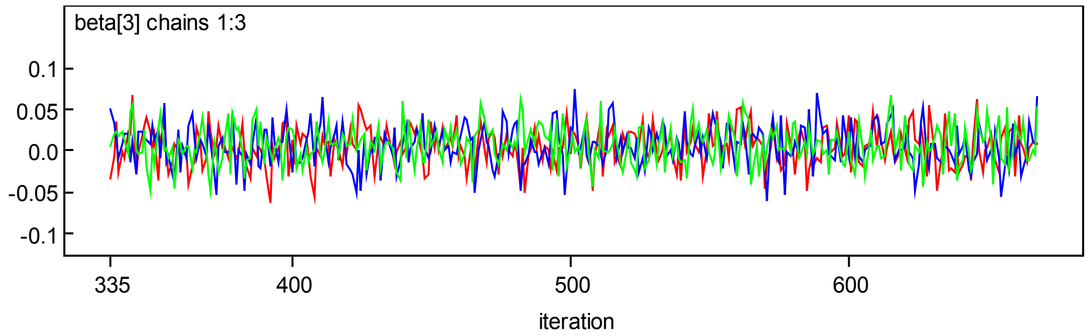

Figure 97. Trace Plot for Kindergarten Full Sample Math Year 05

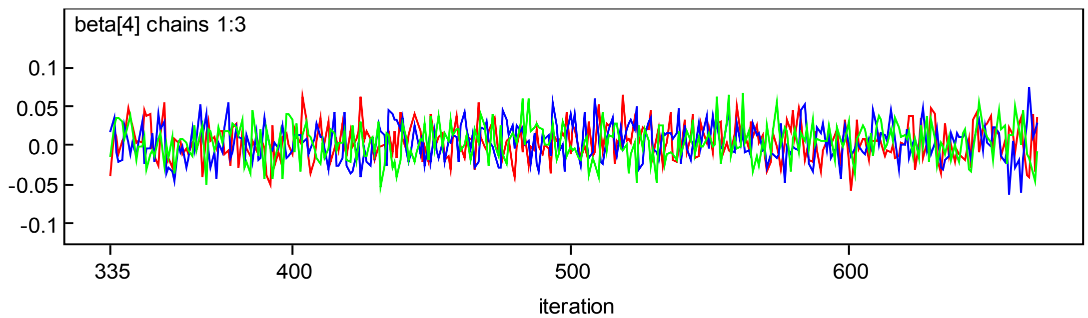

Figure 98. Trace Plot for Kindergarten Full Sample Math Year 06

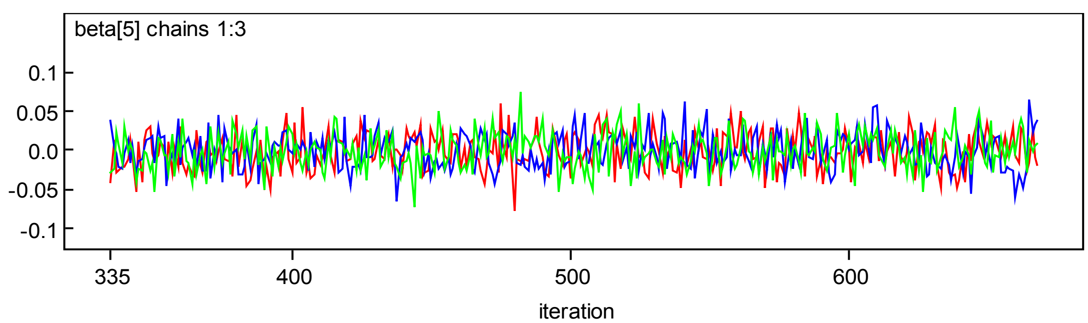

Figure 99. Trace Plot for Kindergarten Full Sample Math Year 07

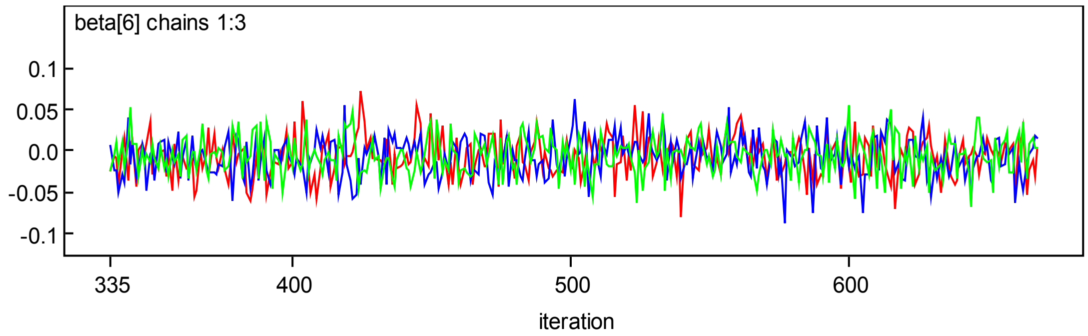

Figure 100. Trace Plot for Kindergarten Full Sample Math Year 08

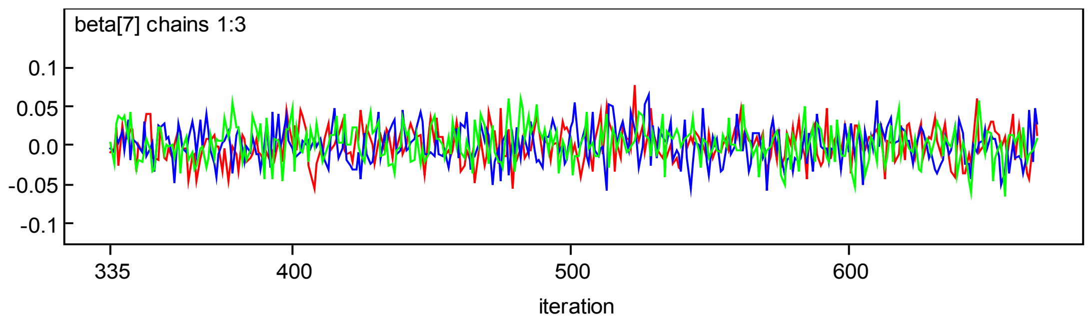

Figure 101. Trace Plot for Kindergarten Full Sample Math Year 09

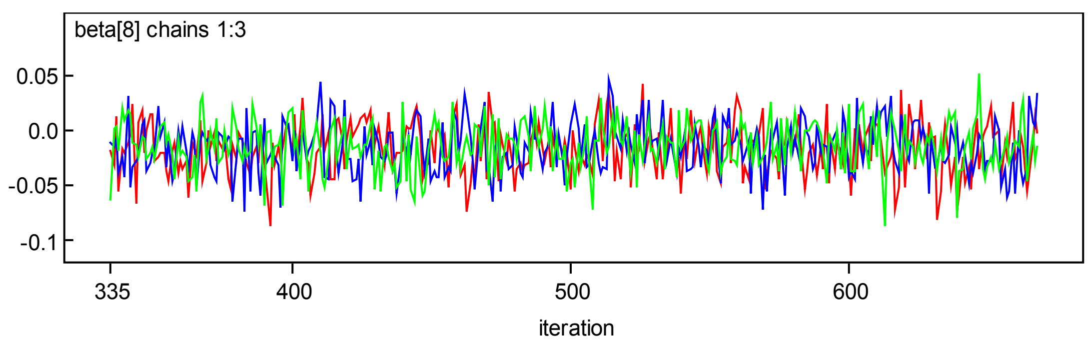

Figure 102. Trace Plot for Kindergarten Full Sample Math Year 10

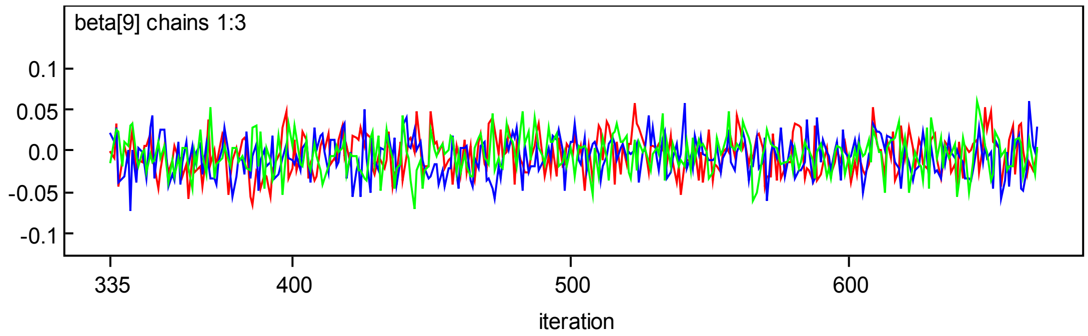

Figure 103. Trace Plot for Kindergarten Full Sample Math Year 11

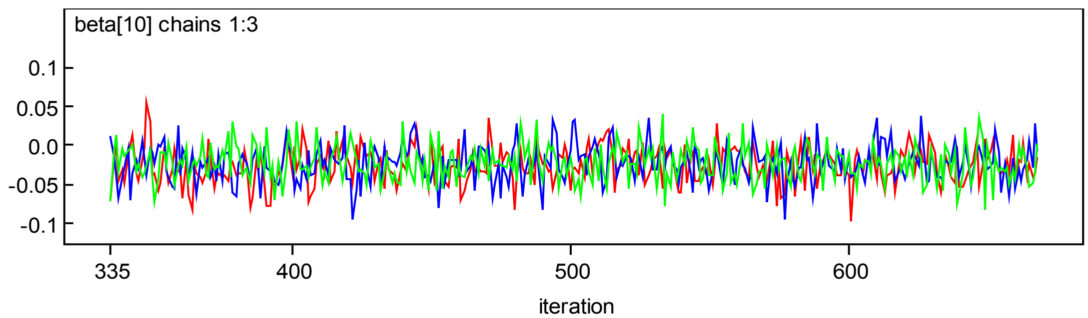

Figure 104. Trace Plot for Kindergarten Full Sample Math Year 12

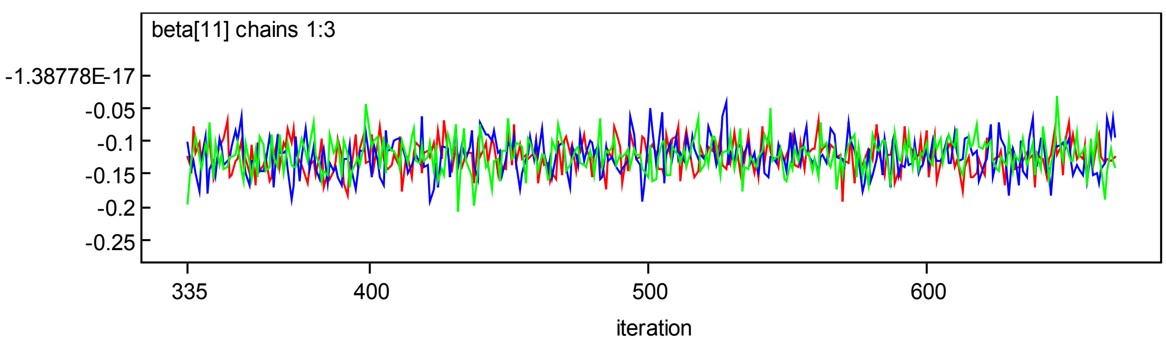

Figure 105. Trace Plot for Kindergarten Full Sample Math Year 13

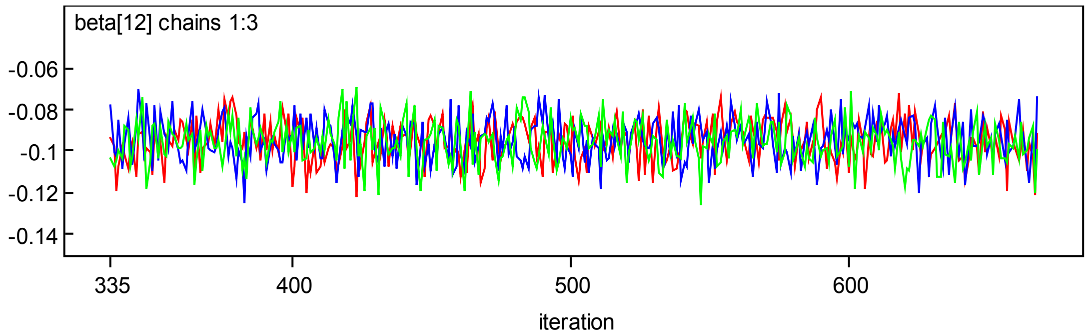

Figure 106. Trace Plot for Kindergarten Full Sample Math Male

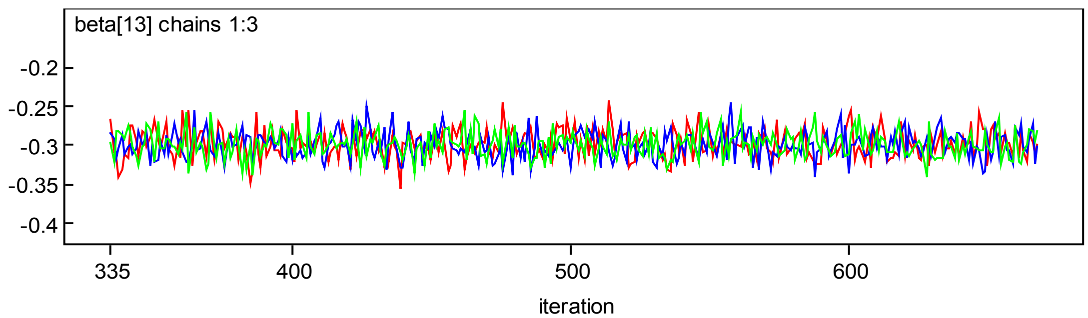

Figure 107. Trace Plot for Kindergarten Full Sample Math Black

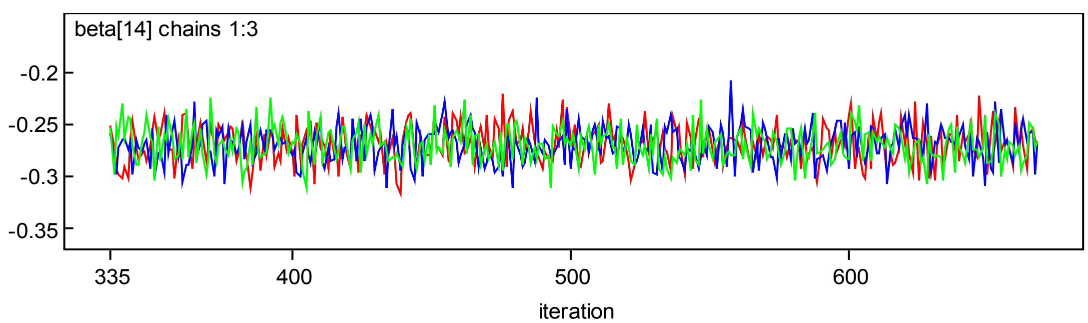

Figure 108. Trace Plot for Kindergarten Full Sample Math Hispanic

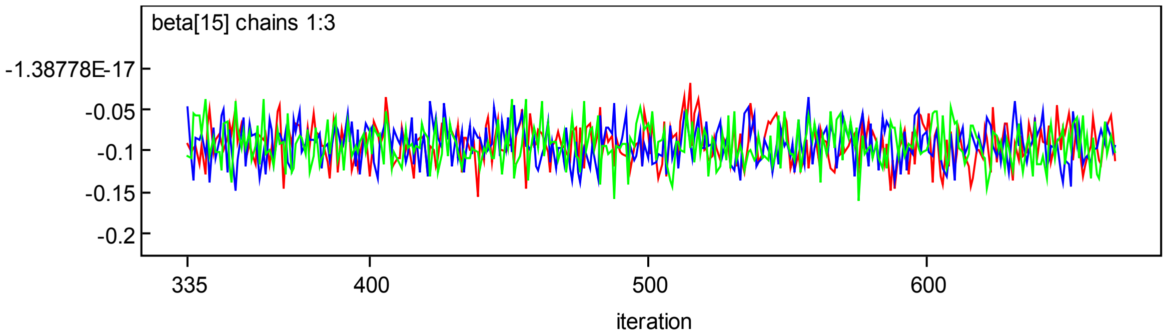

Figure 109. Trace Plot for Kindergarten Full Sample Math Asian

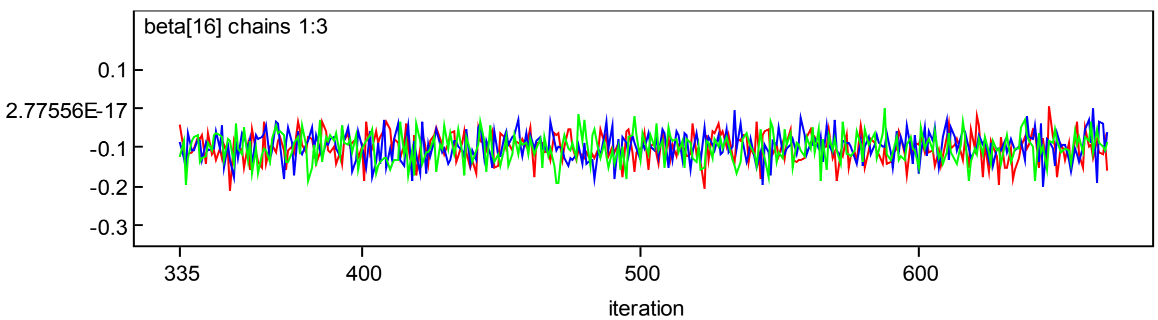

Figure 110. Trace Plot for Kindergarten Full Sample Math Mixed

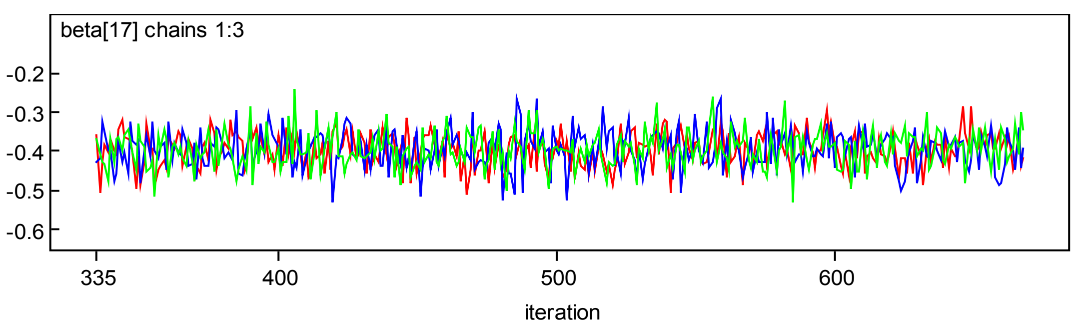

Figure 111. Trace Plot for Kindergarten Full Sample Math Sped 1

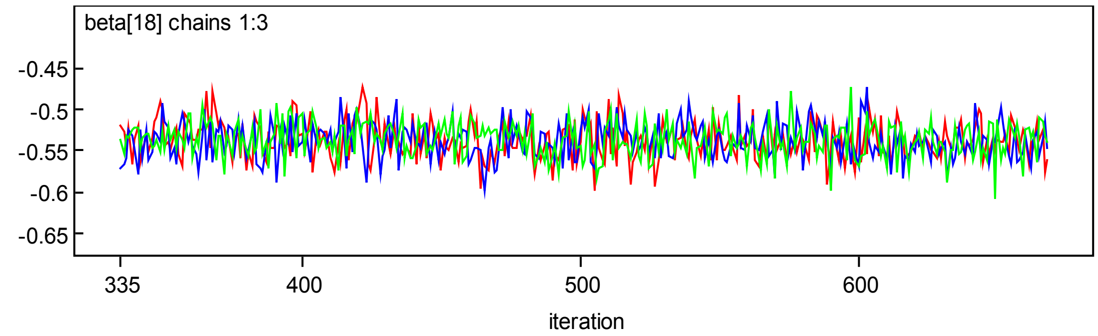

Figure 112. Trace Plot for Kindergarten Full Sample Math Sped 2

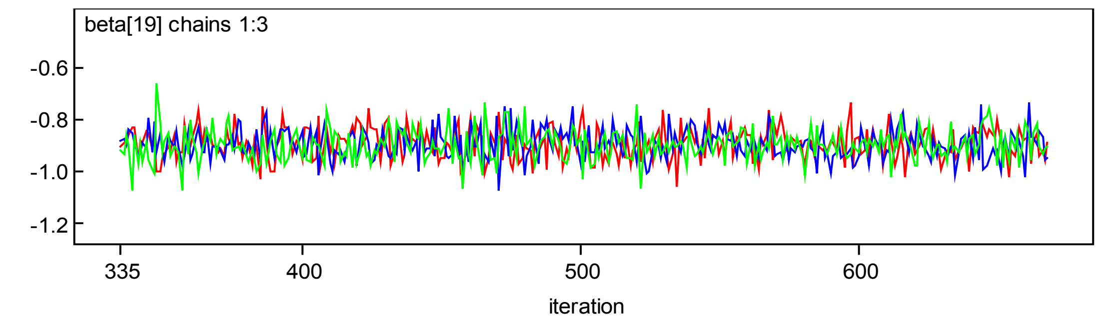

Figure 113. Trace Plot for Kindergarten Full Sample Math Sped 3

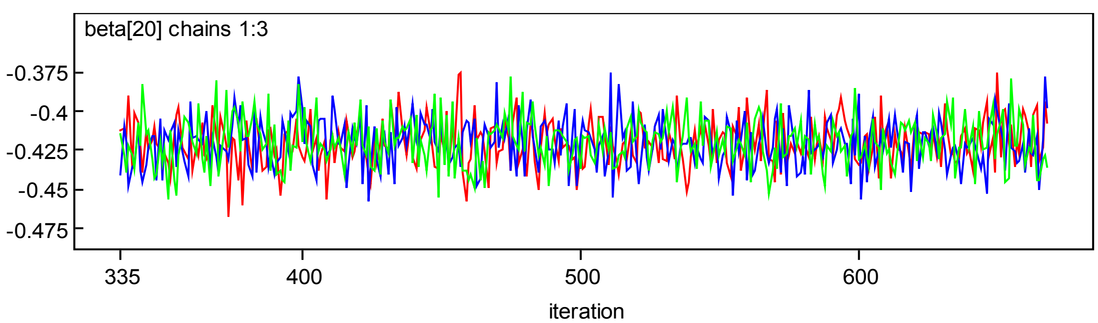

Figure 114. Trace Plot for Kindergarten Full Sample Math Free Lunch

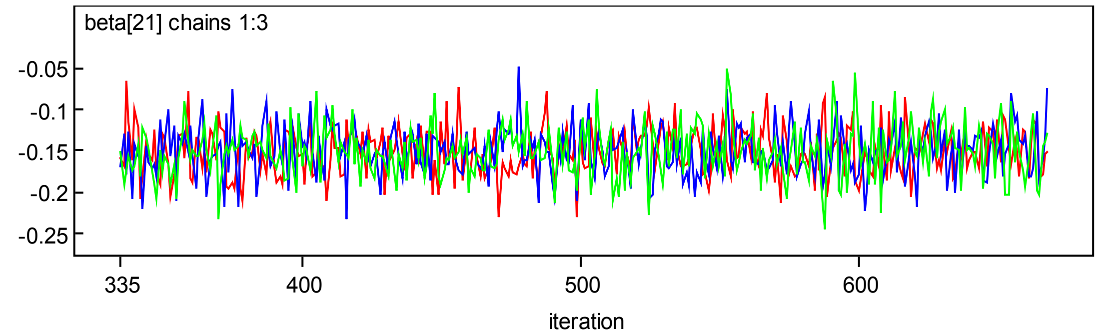

Figure 115. Trace Plot for Kindergarten Full Sample Math Reduced Lunch

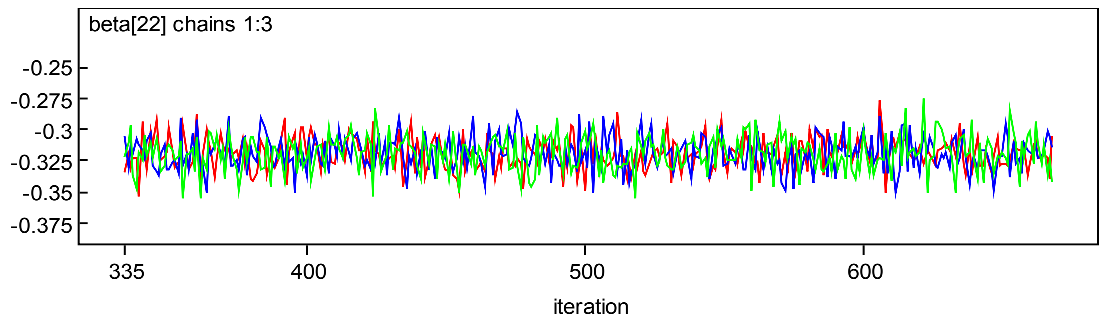

Figure 116. Trace Plot for Kindergarten Full Sample Math Bilingual

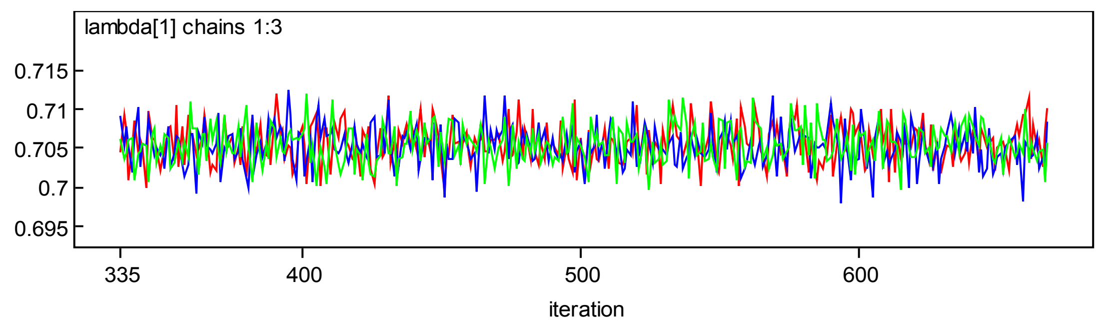

Figure 117. Trace Plot for Kindergarten Full Sample Math Lambda 1

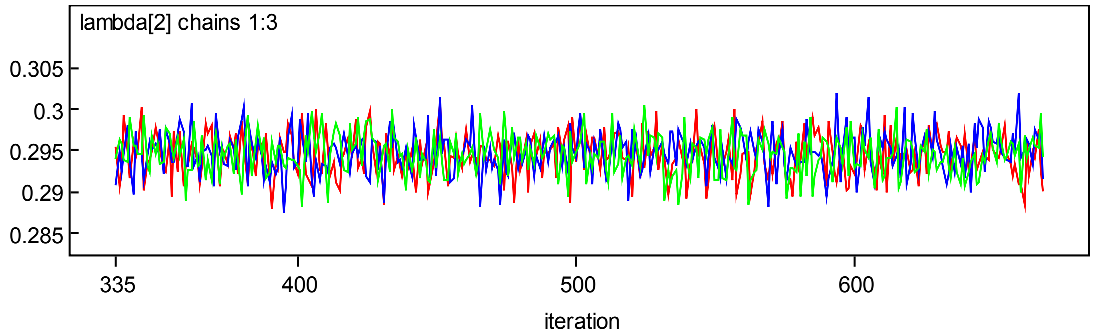

Figure 118. Trace Plot for Kindergarten Full Sample Math Lambda 2

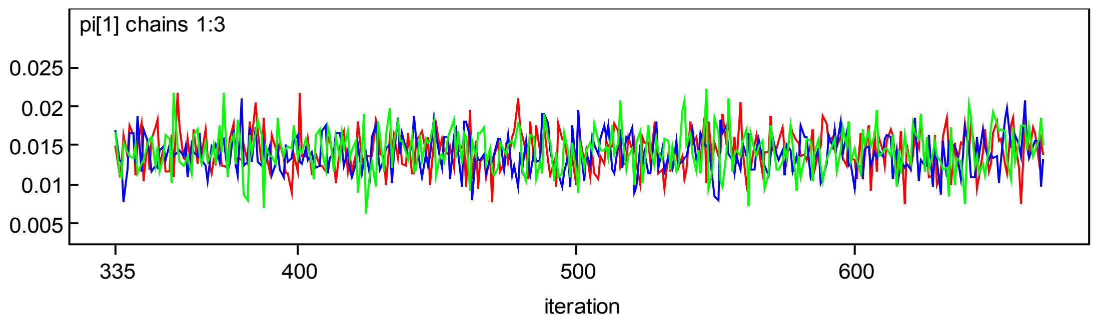

Figure 119. Trace Plot for Kindergarten Full Sample Math Pi 1

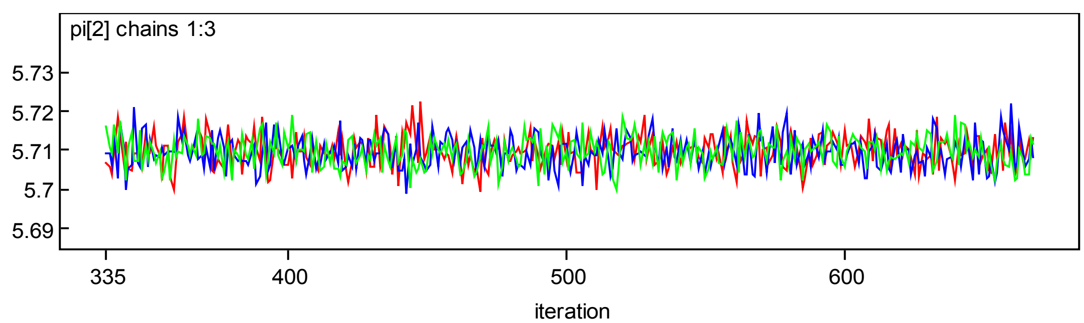

Figure 120. Trace Plot for Kindergarten Full Sample Math Pi 2

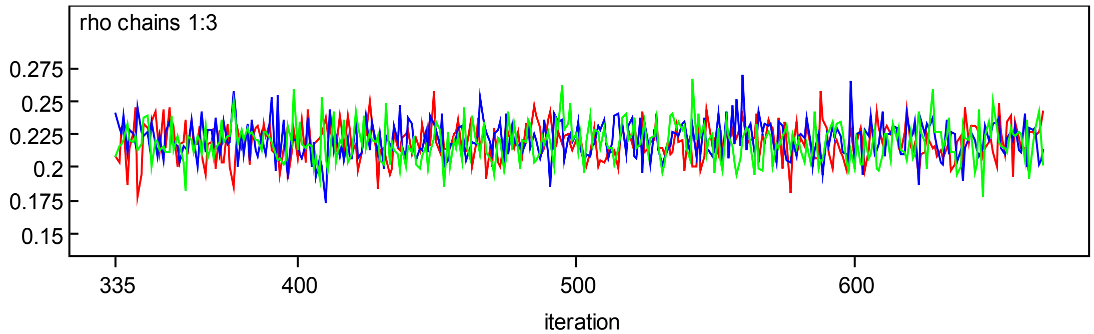

Figure 121. Trace Plot for Kindergarten Full Sample Math Rho

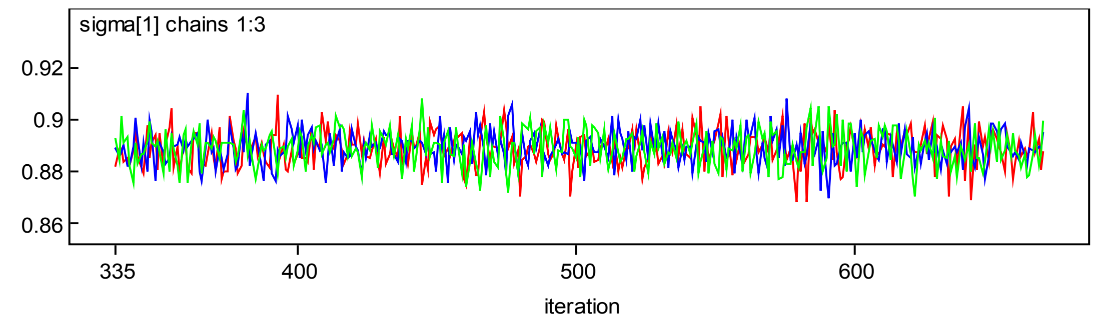

Figure 122. Trace Plot for Kindergarten Full Sample Math Sigma 1

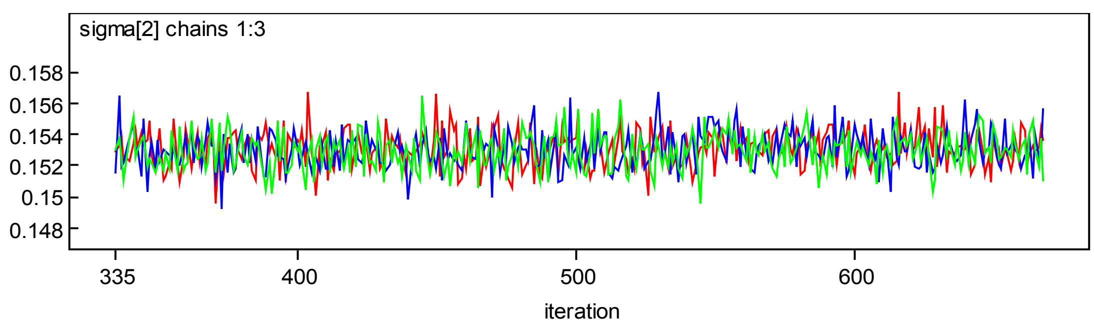

Figure 123. Trace Plot for Kindergarten Full Sample Math Sigma 2

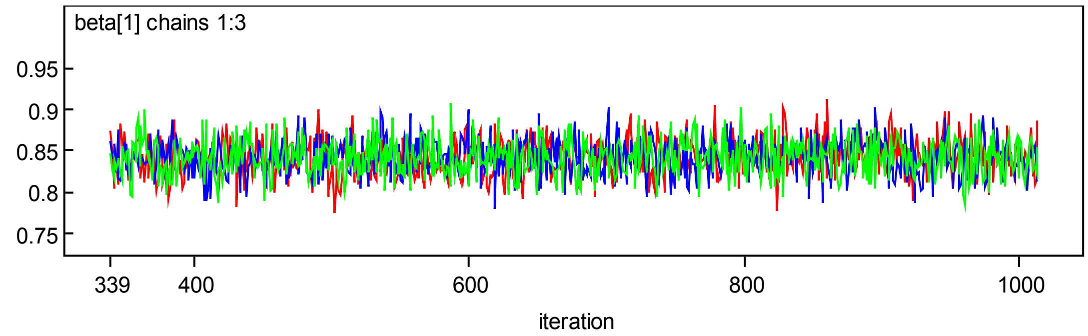

Figure 124. Trace Plot for Kindergarten Full Sample Reading Intercept

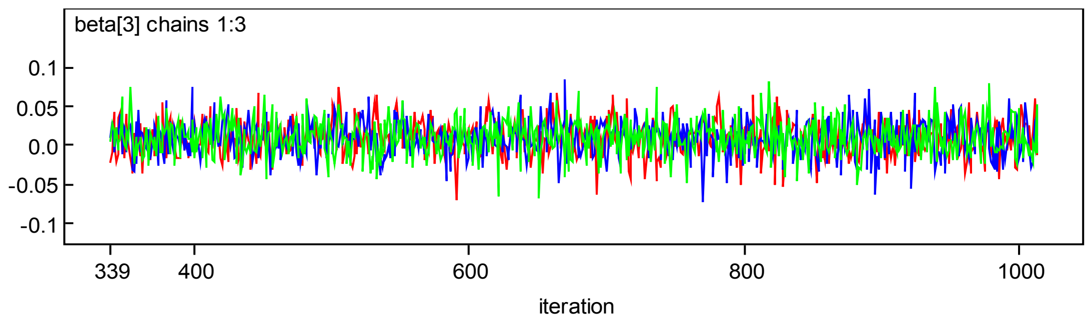

Figure 124. Trace Plot for Kindergarten Full Sample Reading Year 05

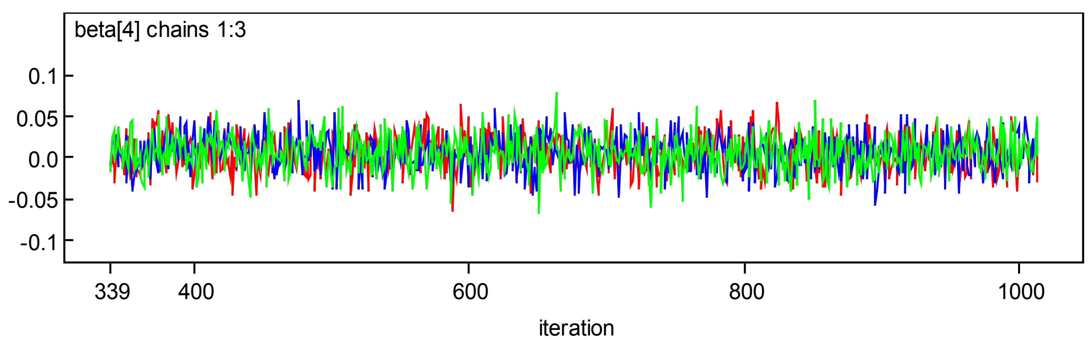

Figure 125. Trace Plot for Kindergarten Full Sample Reading Year 06

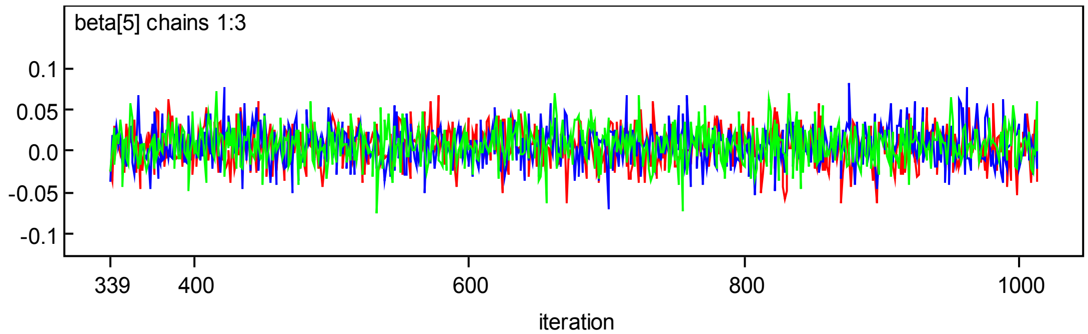

Figure 126. Trace Plot for Kindergarten Full Sample Reading Year 07

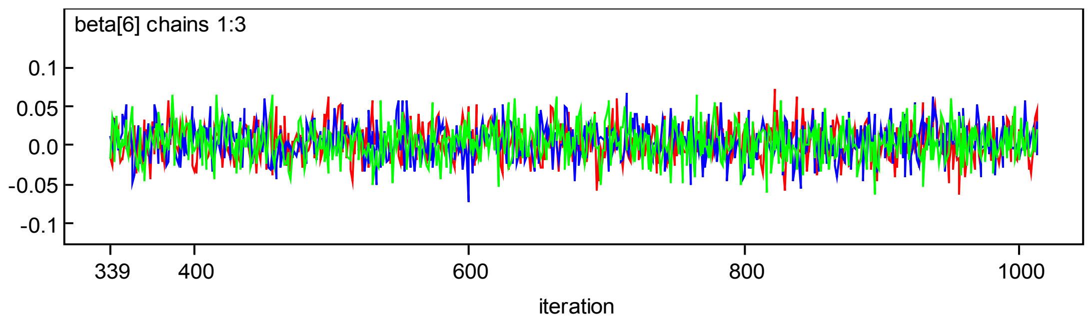

Figure 127. Trace Plot for Kindergarten Full Sample Reading Year 08

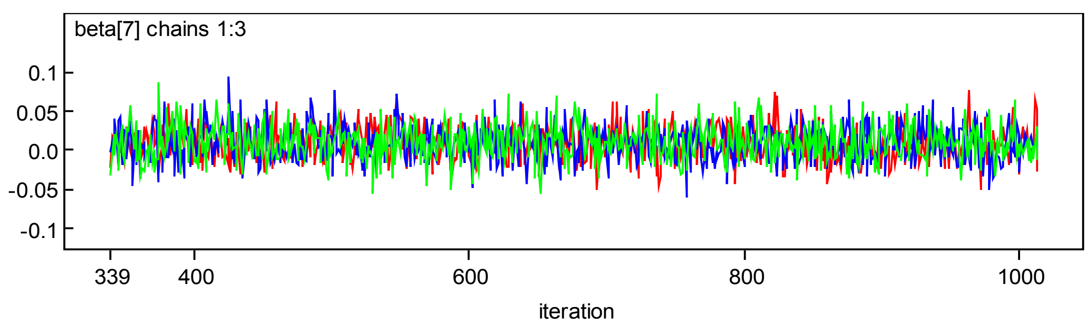

Figure 128. Trace Plot for Kindergarten Full Sample Reading Year 09

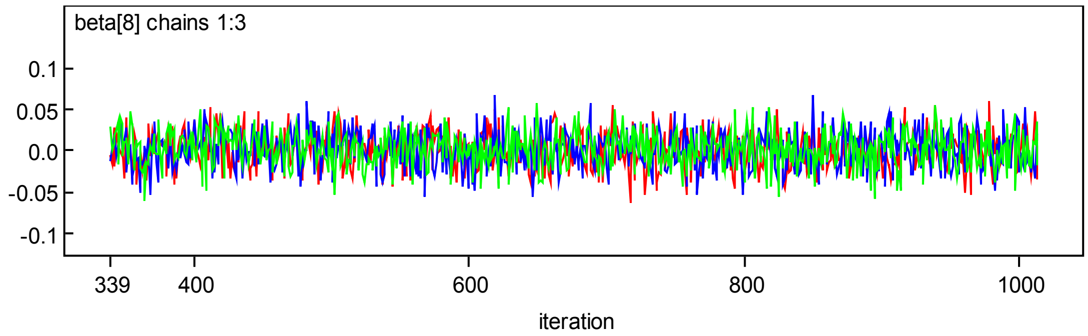

Figure 129. Trace Plot for Kindergarten Full Sample Reading Year 10

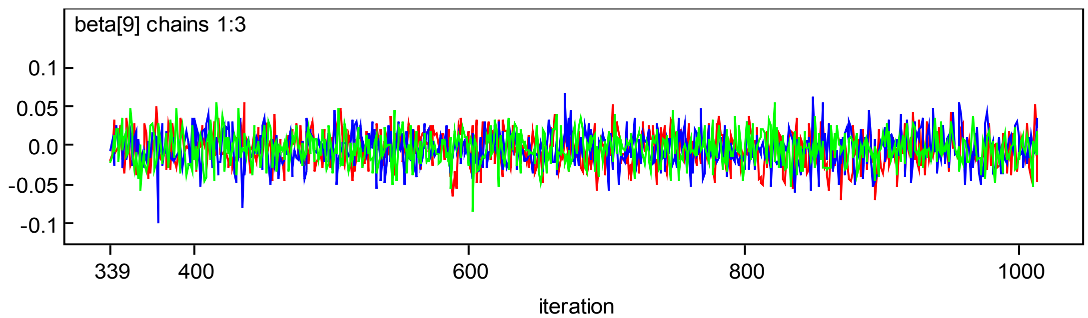

Figure 130. Trace Plot for Kindergarten Full Sample Reading Year 11

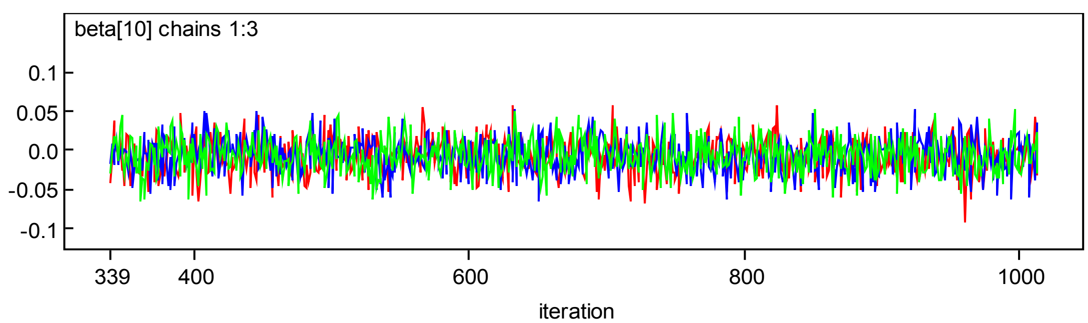

Figure 131. Trace Plot for Kindergarten Full Sample Reading Year 12

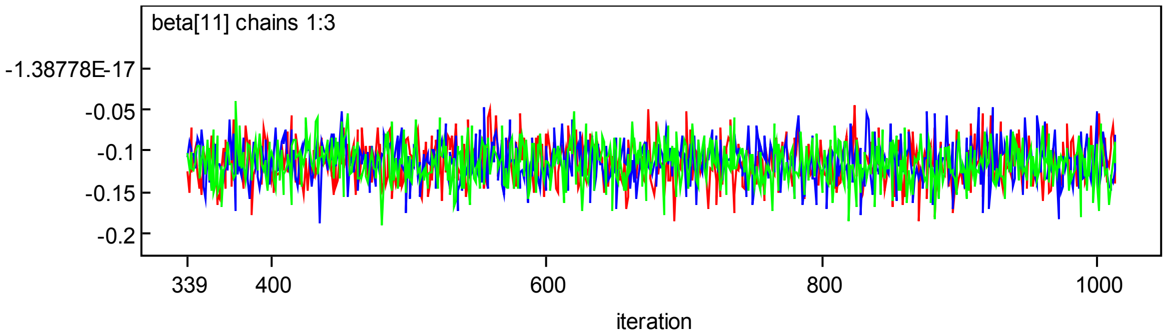

Figure 132. Trace Plot for Kindergarten Full Sample Reading Year 13

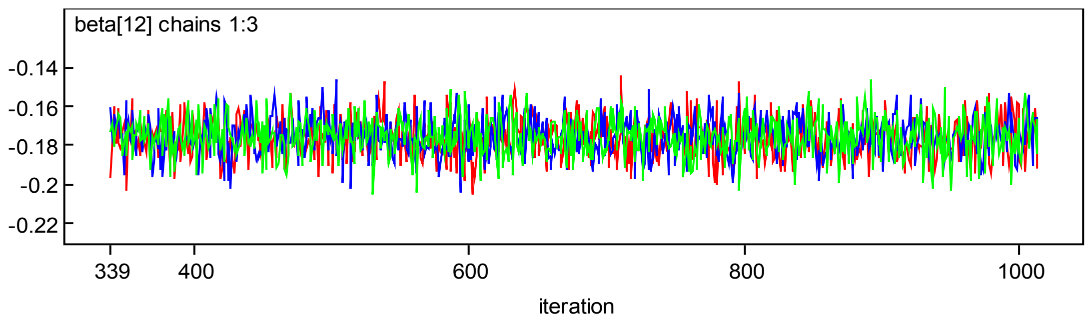

Figure 132. Trace Plot for Kindergarten Full Sample Reading Male

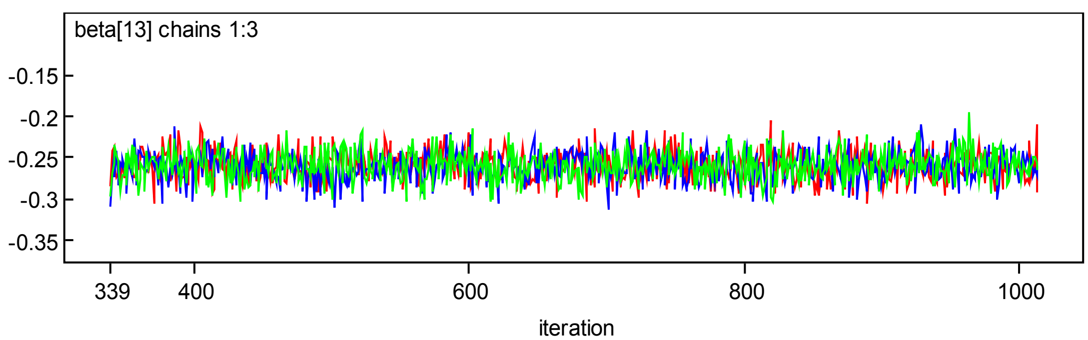

Figure 133. Trace Plot for Kindergarten Full Sample Reading Black

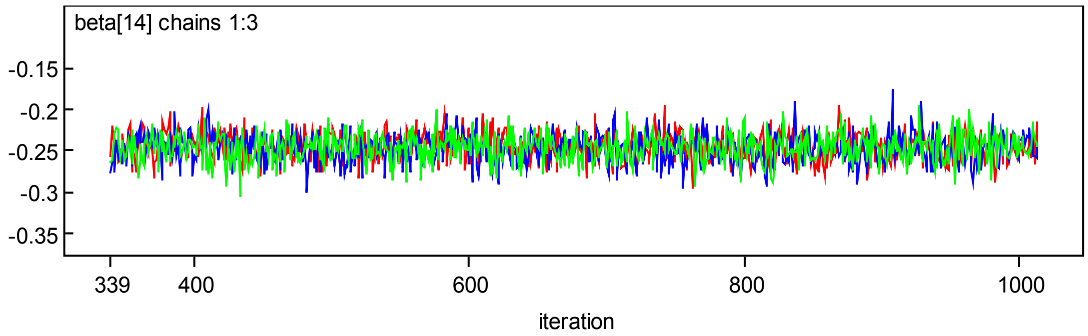

Figure 134. Trace Plot for Kindergarten Full Sample Reading Hispanic

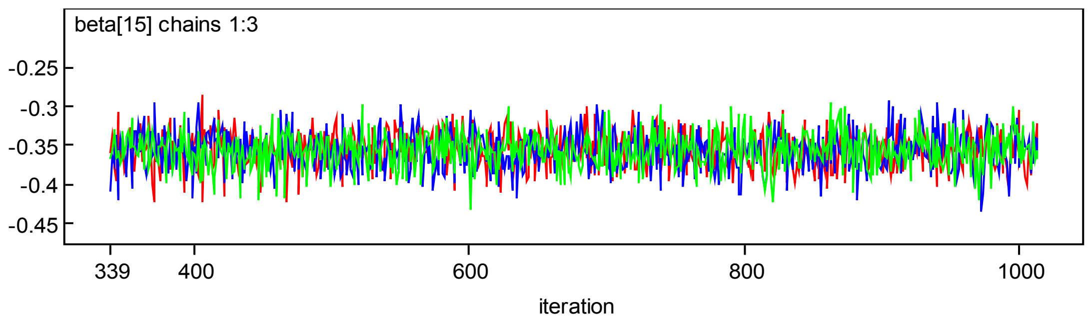

Figure 135. Trace Plot for Kindergarten Full Sample Reading Asian

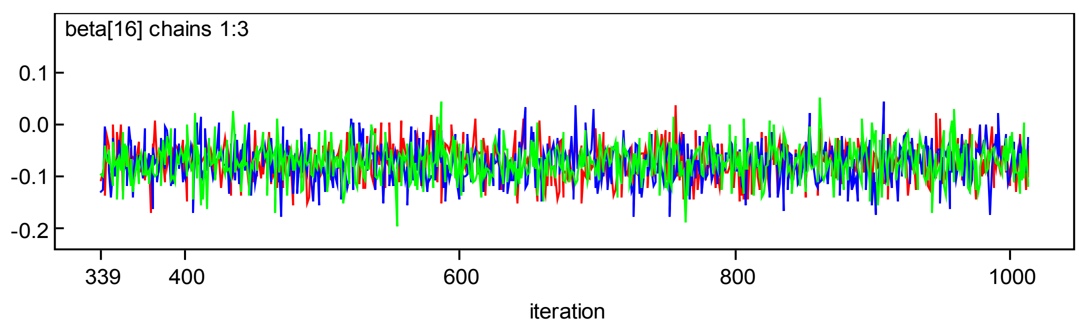

Figure 136. Trace Plot for Kindergarten Full Sample Reading Mixed

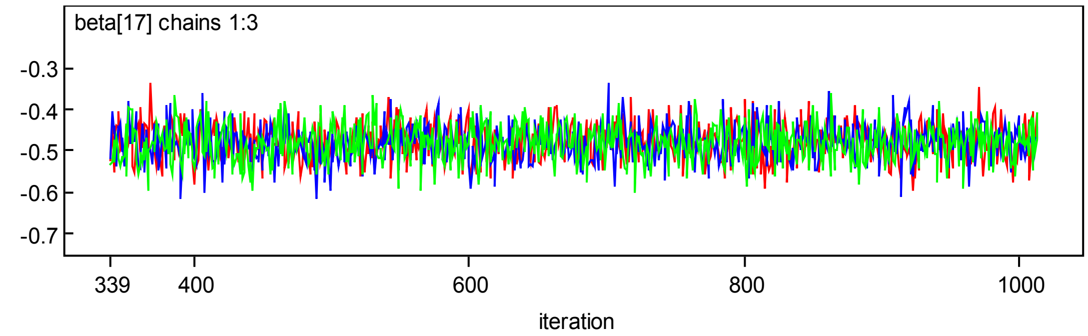

Figure 137. Trace Plot for Kindergarten Full Sample Reading Sped 1

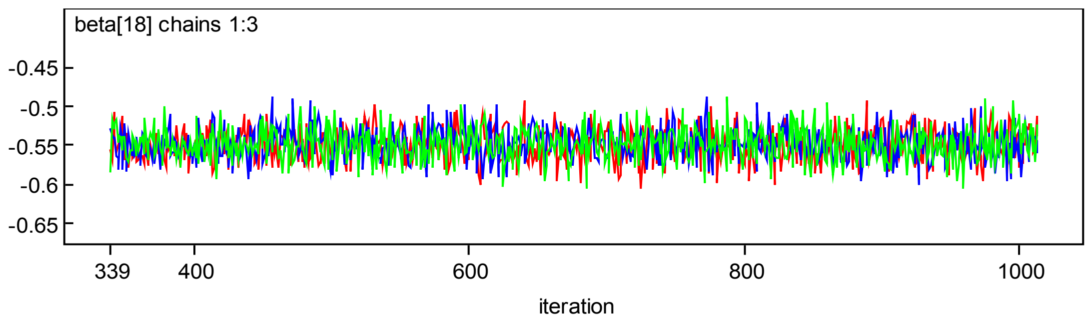

Figure 138. Trace Plot for Kindergarten Full Sample Reading Sped 2

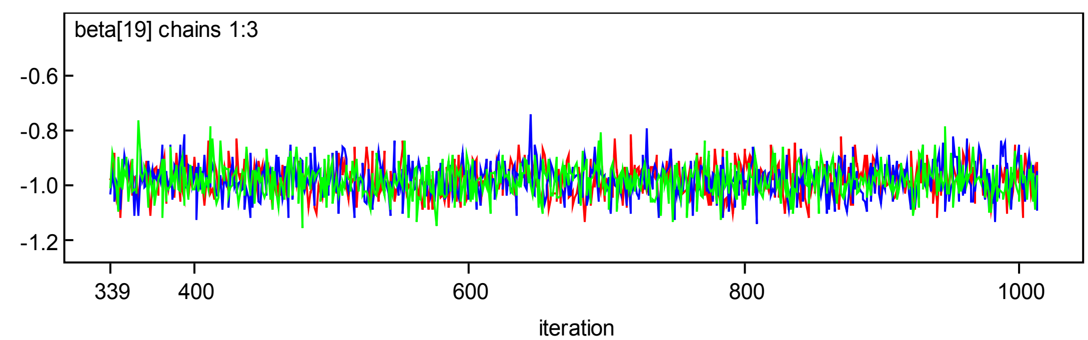

Figure 139. Trace Plot for Kindergarten Full Sample Reading Sped 3

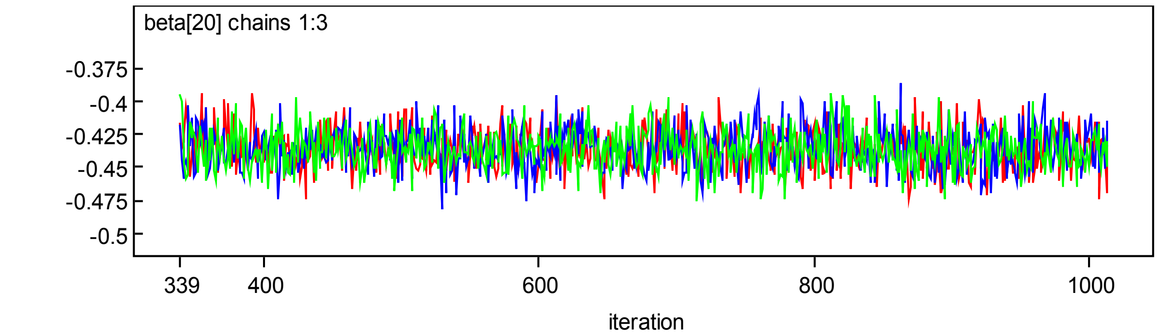

Figure 140. Trace Plot for Kindergarten Full Sample Reading Free Lunch

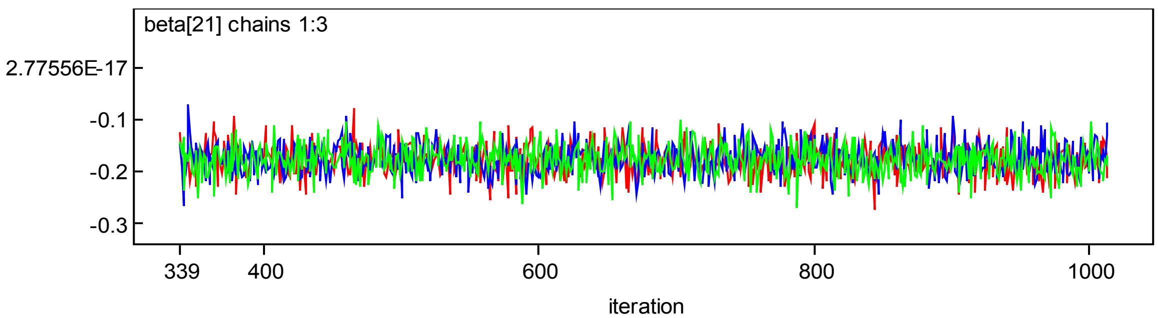

Figure 141. Trace Plot for Kindergarten Full Sample Reading Reduced Lunch

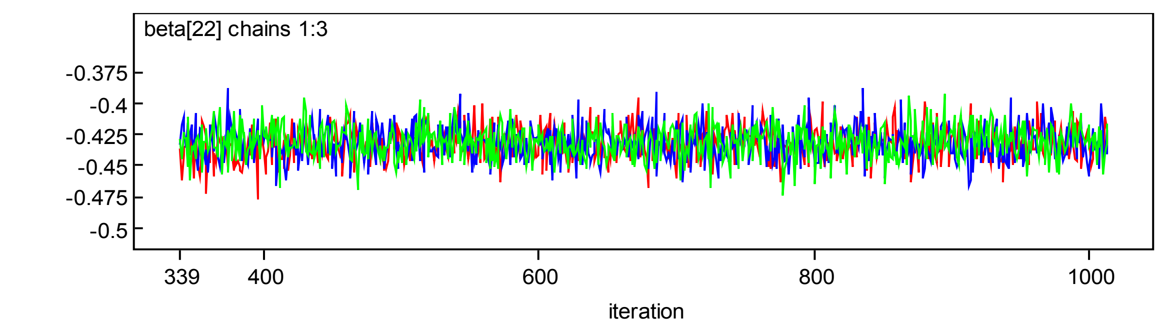

Figure 142. Trace Plot for Kindergarten Full Sample Reading Bilingual

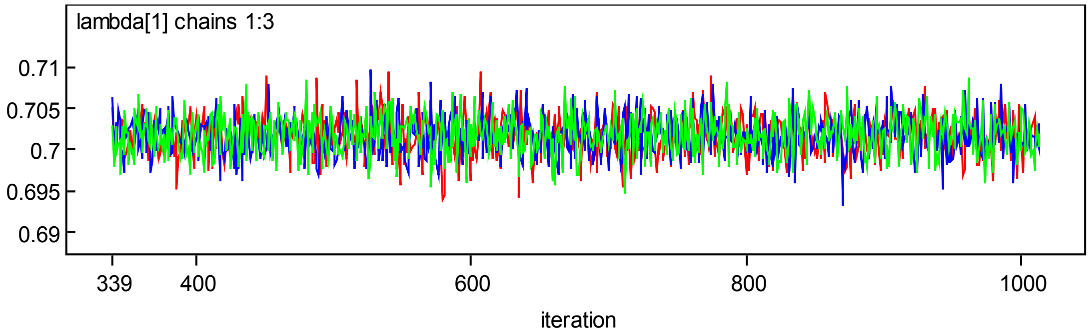

Figure 143. Trace Plot for Kindergarten Full Sample Reading Lambda 1

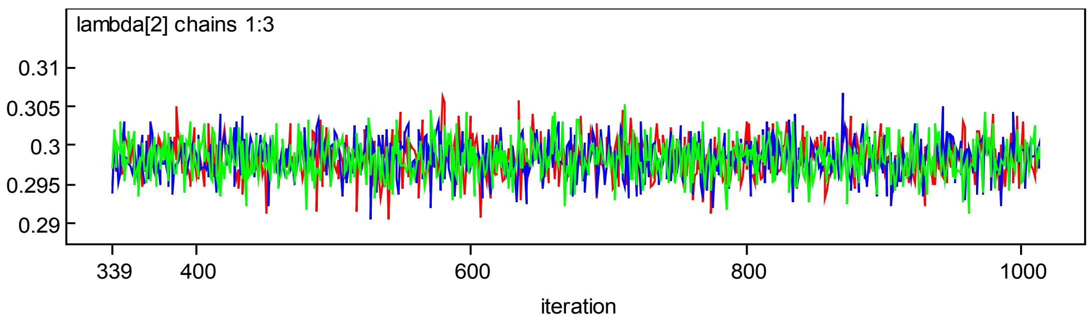

Figure 144. Trace Plot for Kindergarten Full Sample Reading Lambda 2

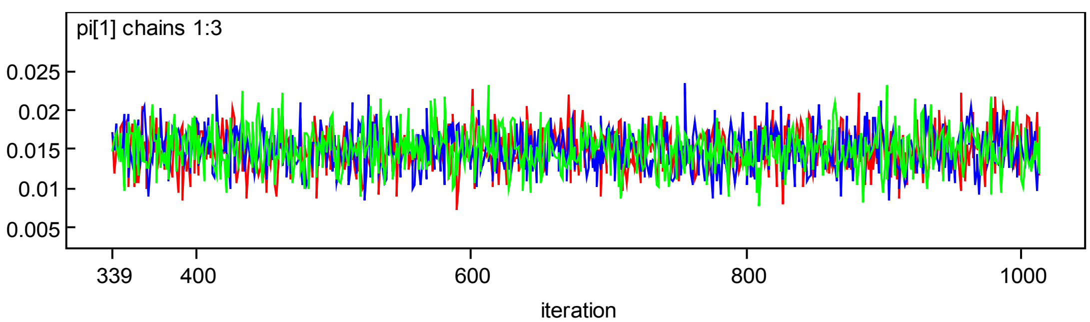

Figure 145. Trace Plot for Kindergarten Full Sample Reading Pi 1

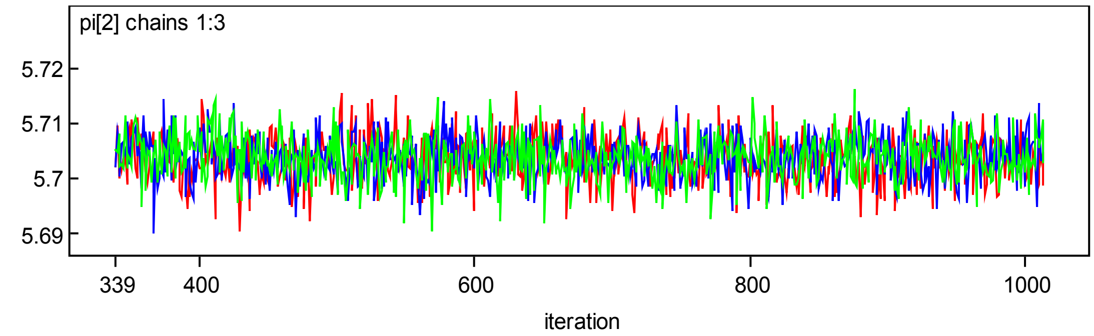

Figure 146. Trace Plot for Kindergarten Full Sample Reading Pi 2

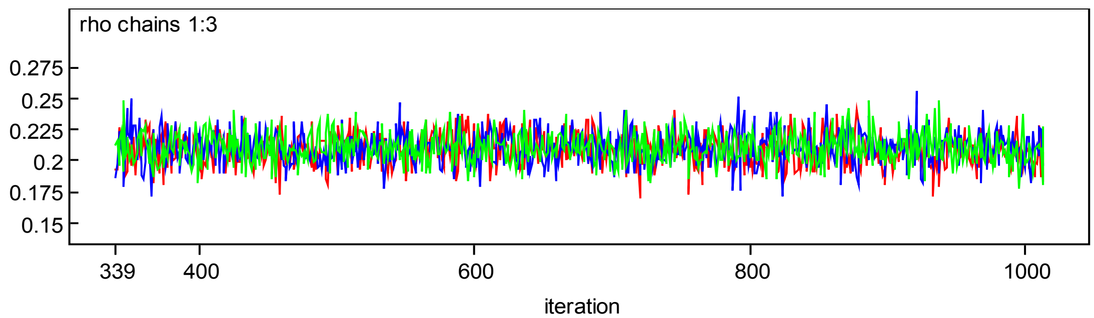

Figure 147. Trace Plot for Kindergarten Full Sample Reading Rho

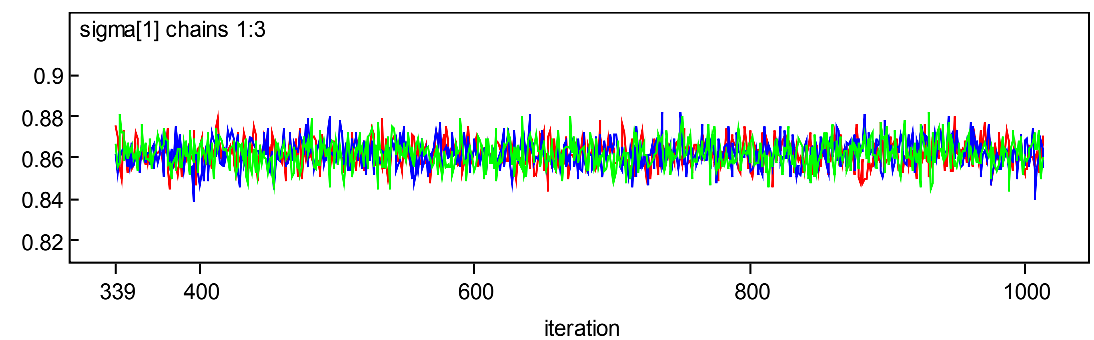

Figure 148. Trace Plot for Kindergarten Full Sample Reading Sigma 1

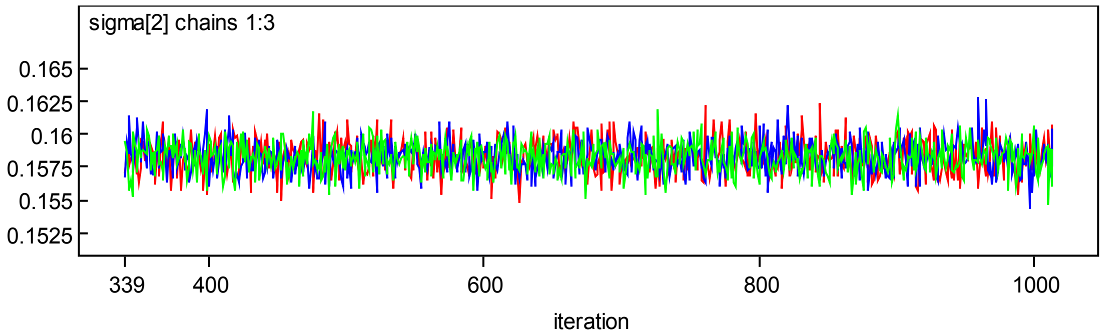

Figure 149. Trace Plot for Kindergarten Full Sample Reading Sigma 2

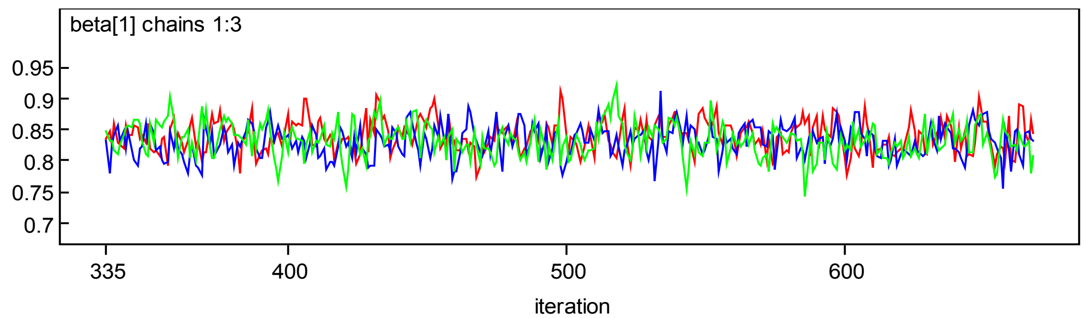

Figure 150. Trace Plot for Grade 3 Full Sample Math Intercept

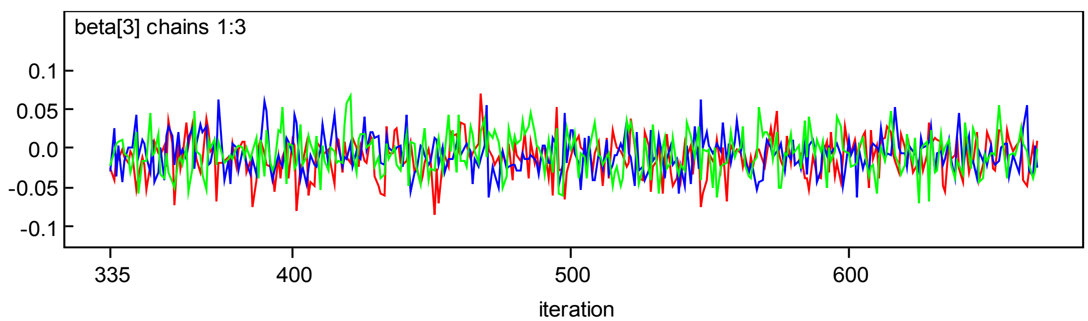

Figure 151. Trace Plot for Grade 3 Full Sample Math Year 06

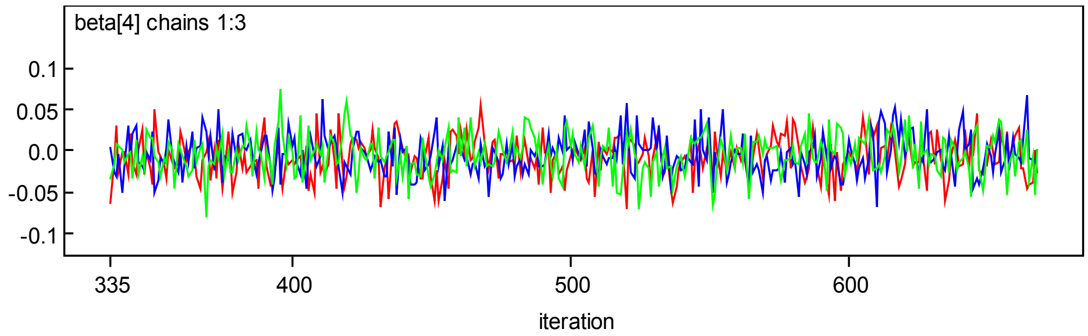

Figure 152. Trace Plot for Grade 3 Full Sample Math Year 07

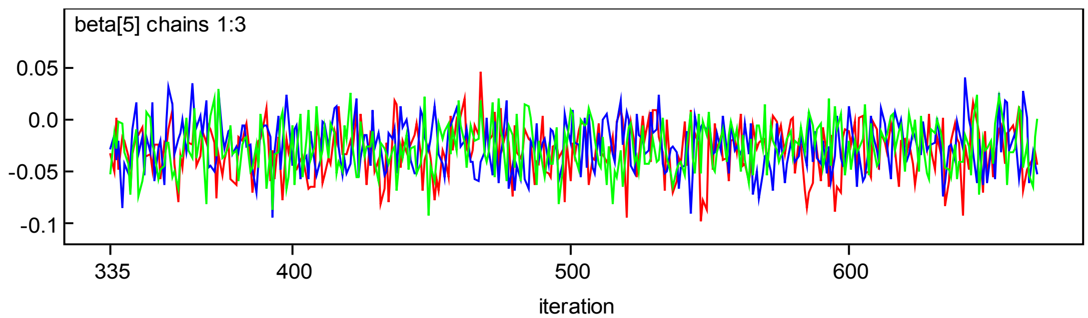

Figure 153. Trace Plot for Grade 3 Full Sample Math Year 08

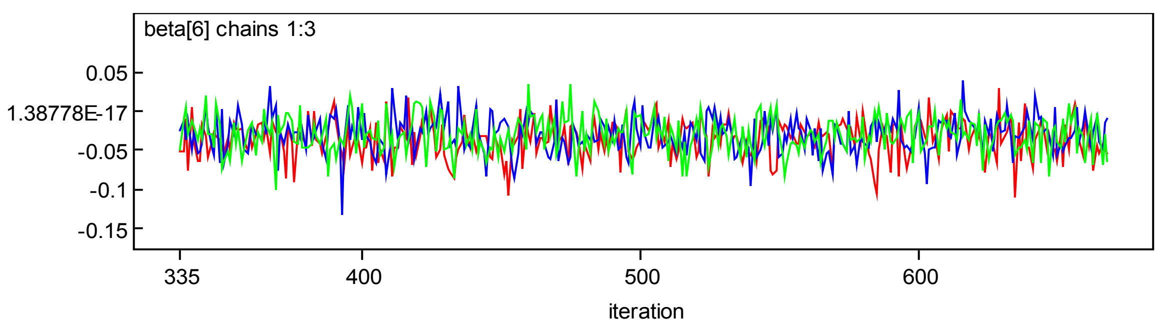

Figure 154. Trace Plot for Grade 3 Full Sample Math Year 09

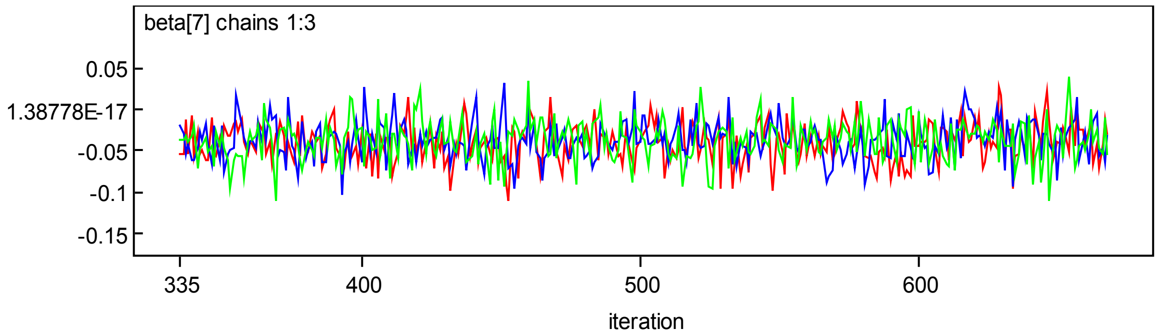

Figure 155. Trace Plot for Grade 3 Full Sample Math Year 10

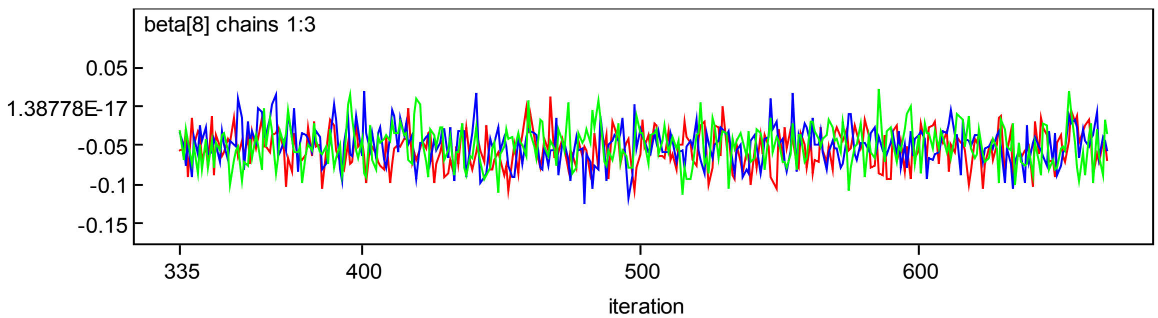

Figure 156. Trace Plot for Grade 3 Full Sample Math Year 11

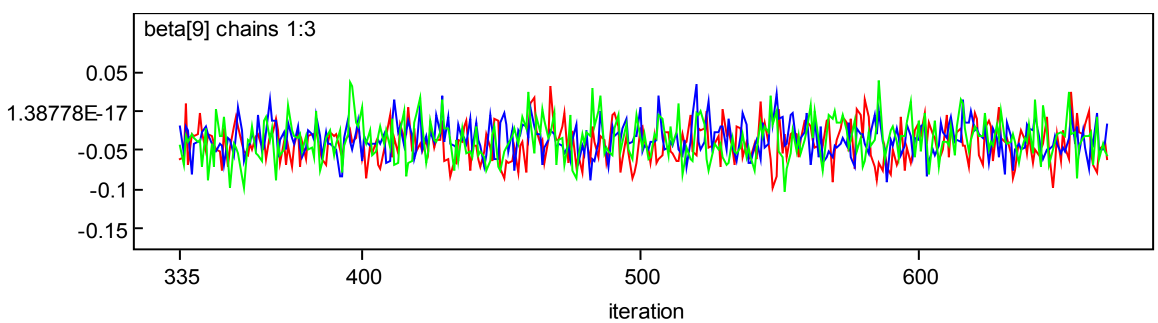

Figure 157. Trace Plot for Grade 3 Full Sample Math Year 12

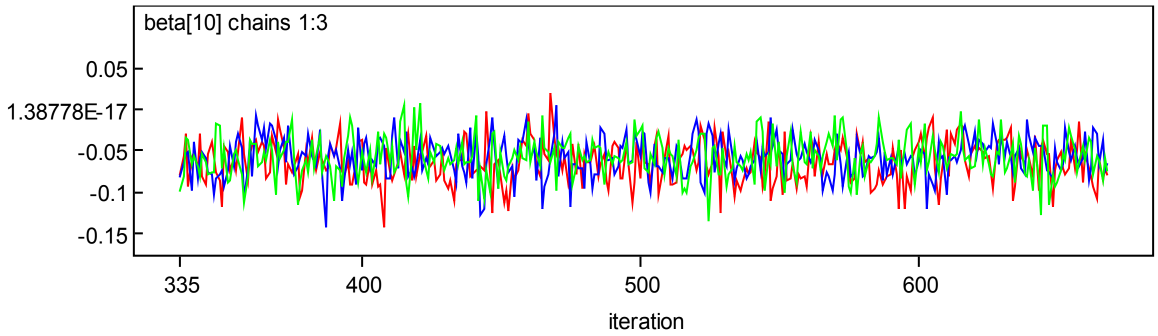

Figure 158. Trace Plot for Grade 3 Full Sample Math Year 13

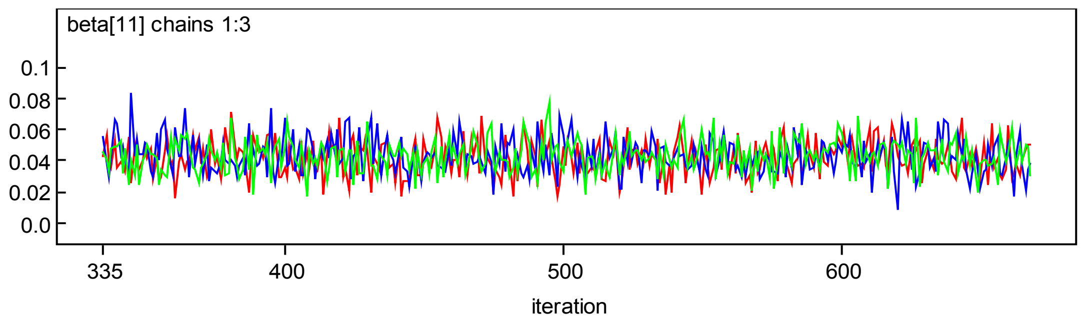

Figure 159. Trace Plot for Grade 3 Full Sample Math Male

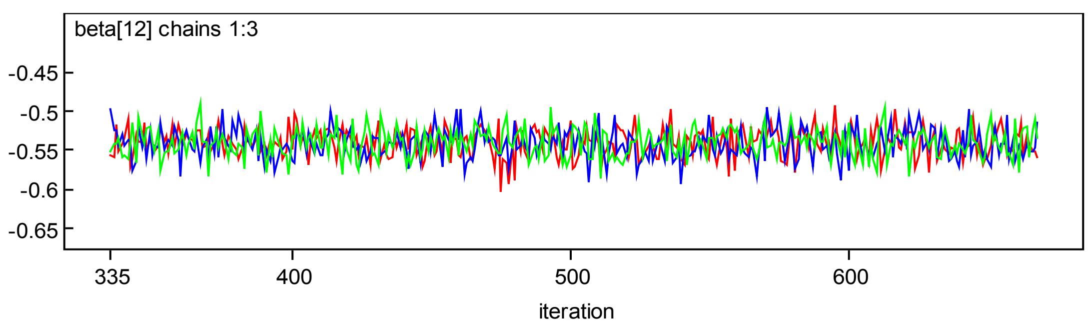

Figure 160. Trace Plot for Grade 3 Full Sample Math Black

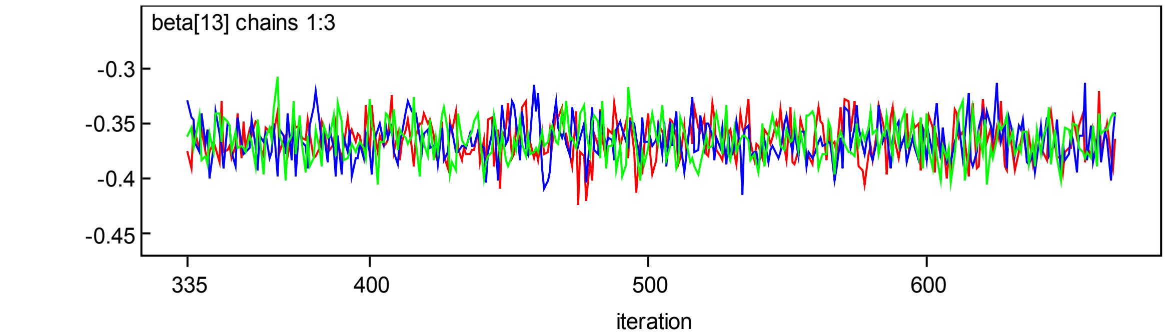

Figure 161. Trace Plot for Grade 3 Full Sample Math Hispanic

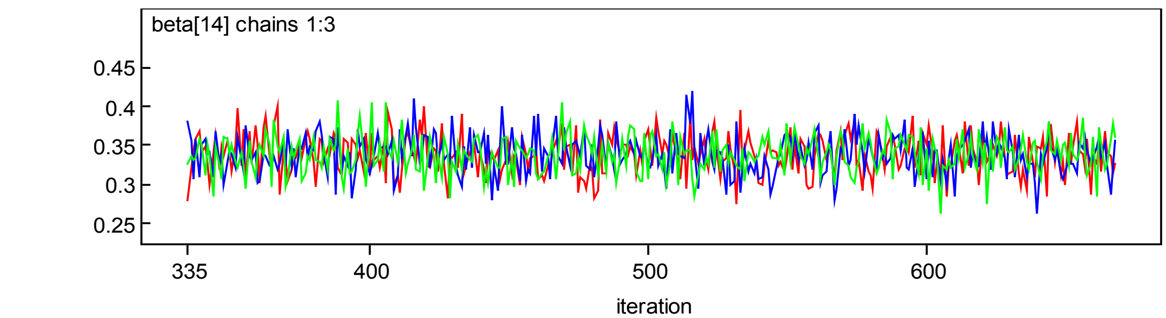

Figure 162. Trace Plot for Grade 3 Full Sample Math Asian

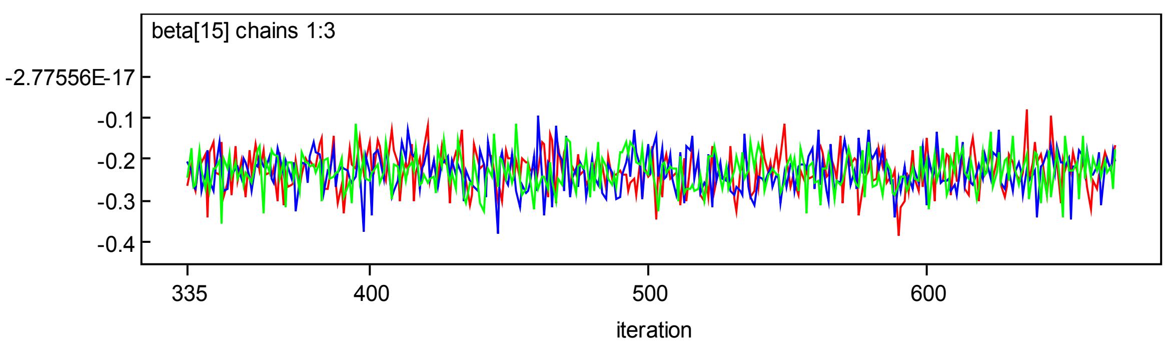

Figure 163. Trace Plot for Grade 3 Full Sample Math Mixed

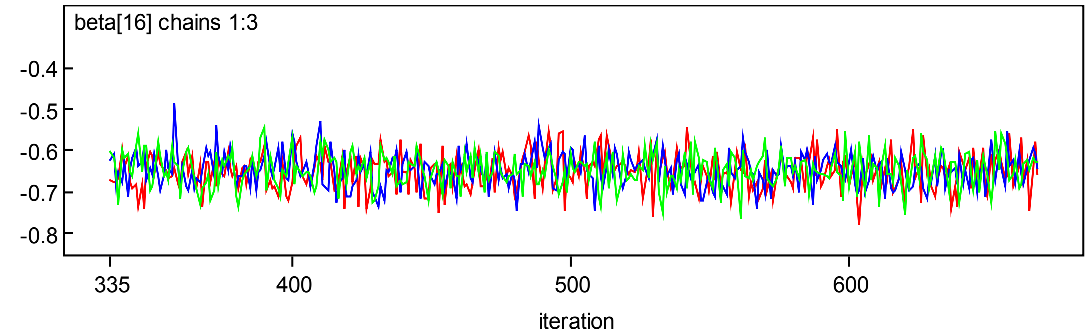

Figure 164. Trace Plot for Grade 3 Full Sample Math Sped 1

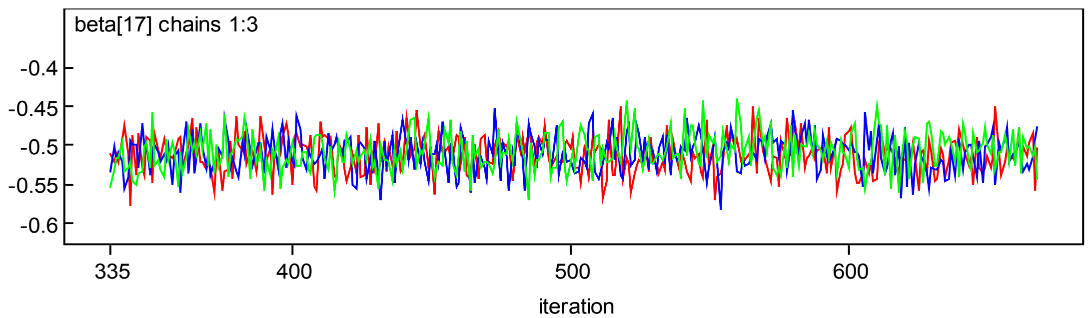

Figure 165. Trace Plot for Grade 3 Full Sample Math Sped 2

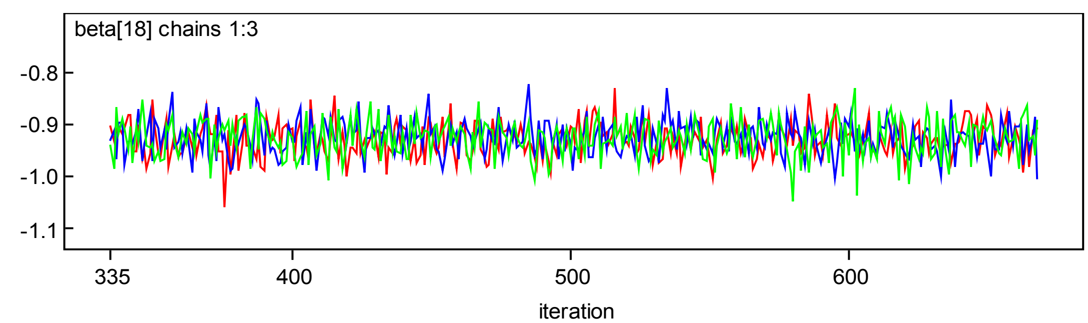

Figure 166. Trace Plot for Grade 3 Full Sample Math Sped 3

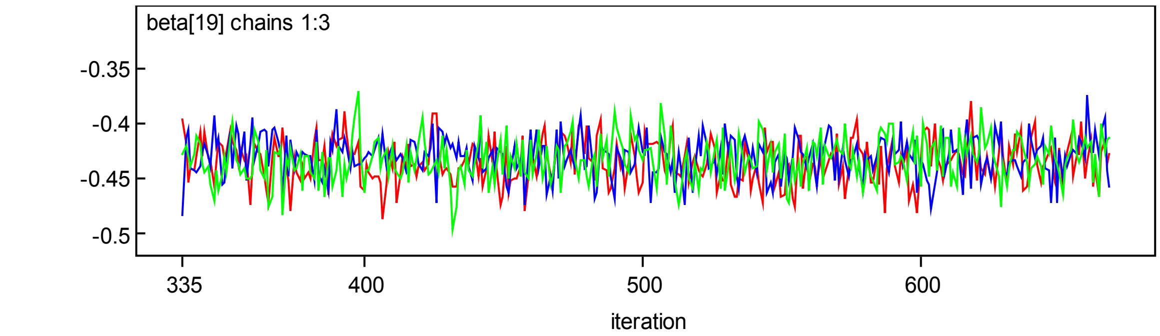

Figure 167. Trace Plot for Grade 3 Full Sample Math Free Lunch

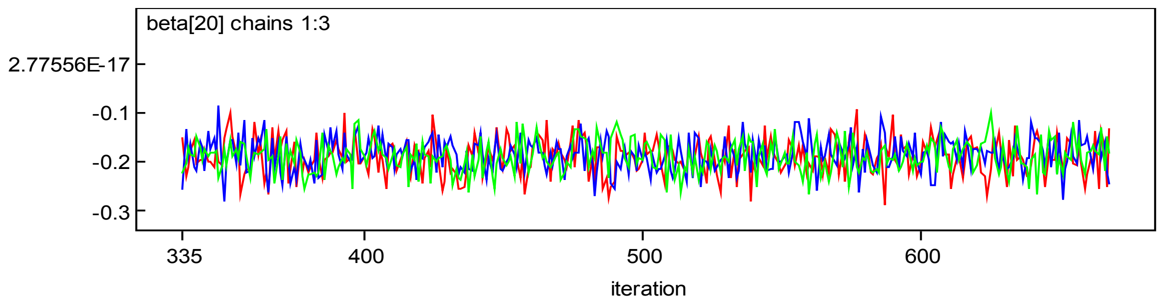

Figure 168. Trace Plot for Grade 3 Full Sample Math Reduced Lunch

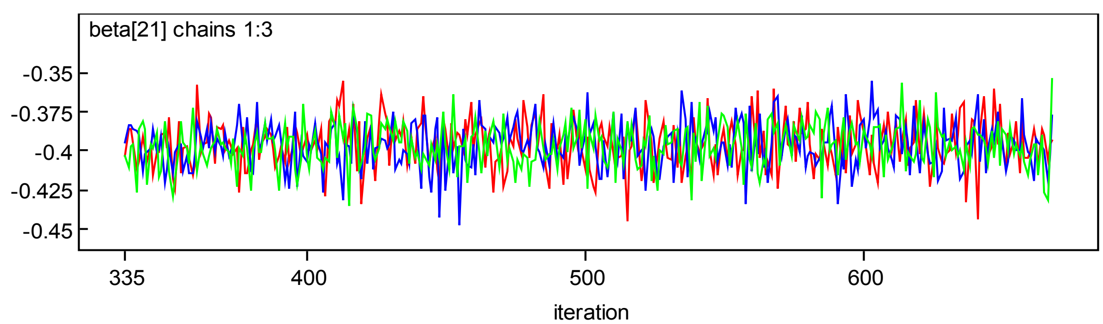

Figure 169. Trace Plot for Grade 3 Full Sample Math Bilingual

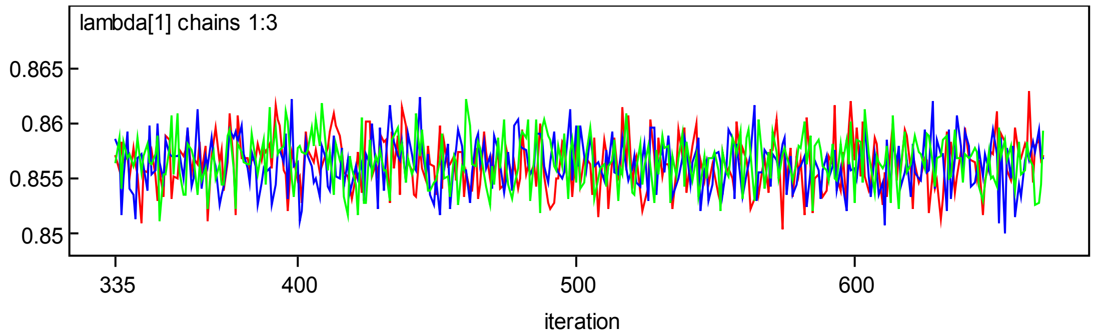

Figure 170. Trace Plot for Grade 3 Full Sample Math Lambda 1

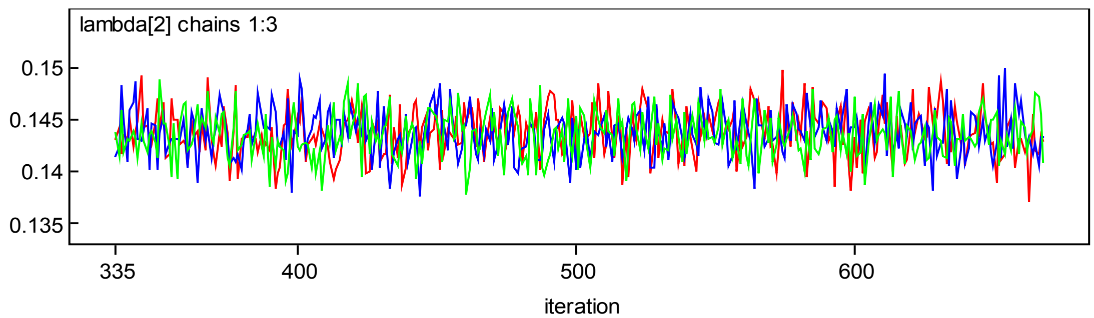

Figure 171. Trace Plot for Grade 3 Full Sample Math Lambda 2

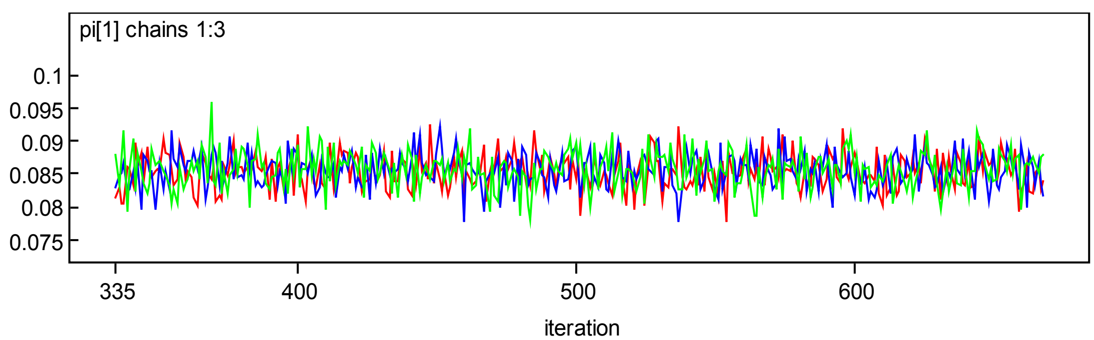

Figure 172. Trace Plot for Grade 3 Full Sample Math Pi 1

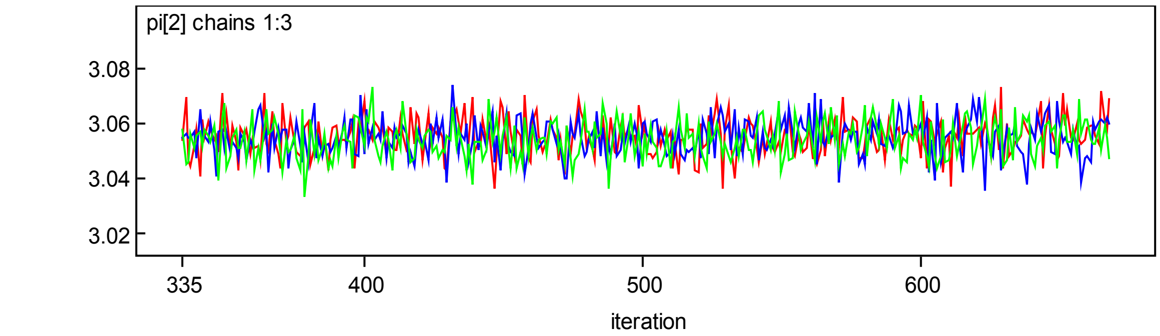

Figure 173. Trace Plot for Grade 3 Full Sample Math Pi 2

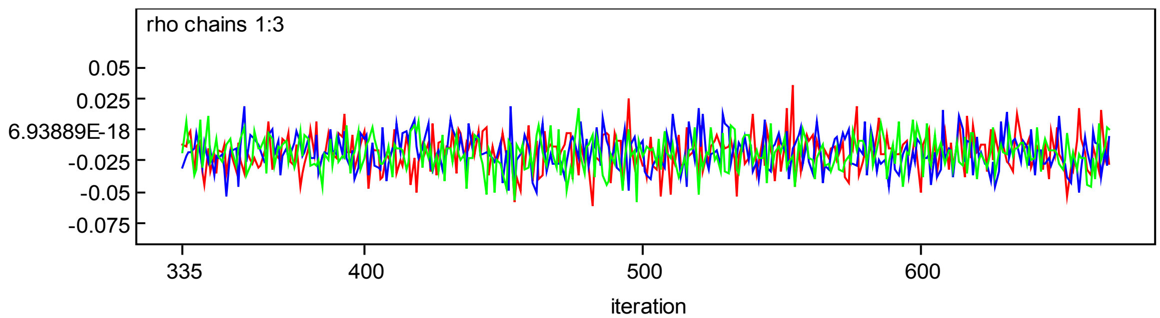

Figure 174. Trace Plot for Grade 3 Full Sample Math Rho

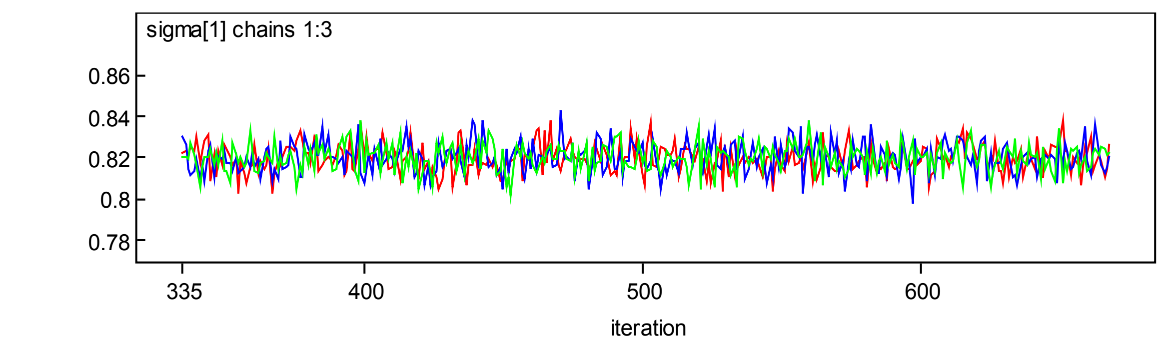

Figure 175. Trace Plot for Grade 3 Full Sample Math Sigma 1

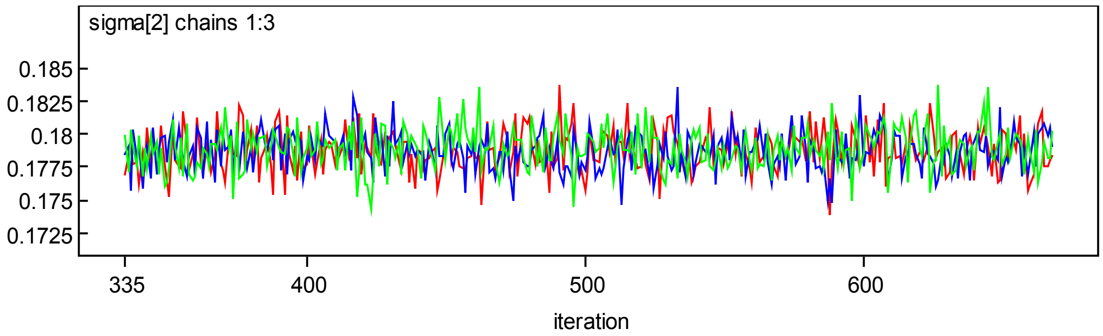

Figure 176. Trace Plot for Grade 3 Full Sample Math Sigma 2

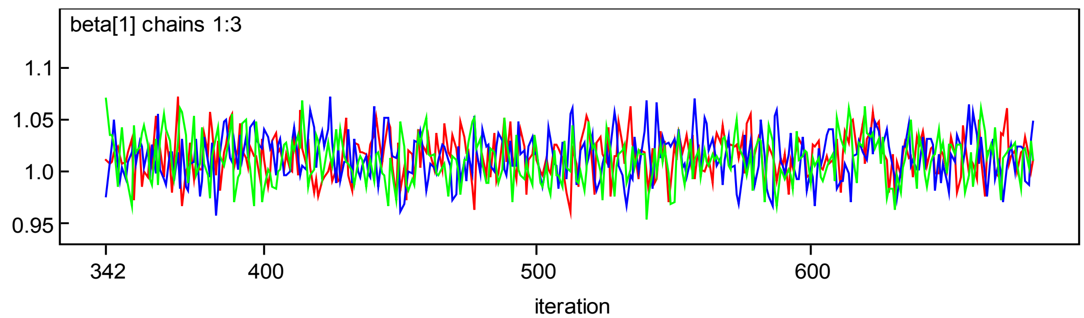

Figure 177. Trace Plot for Grade 3 Full Sample ELA Intercept

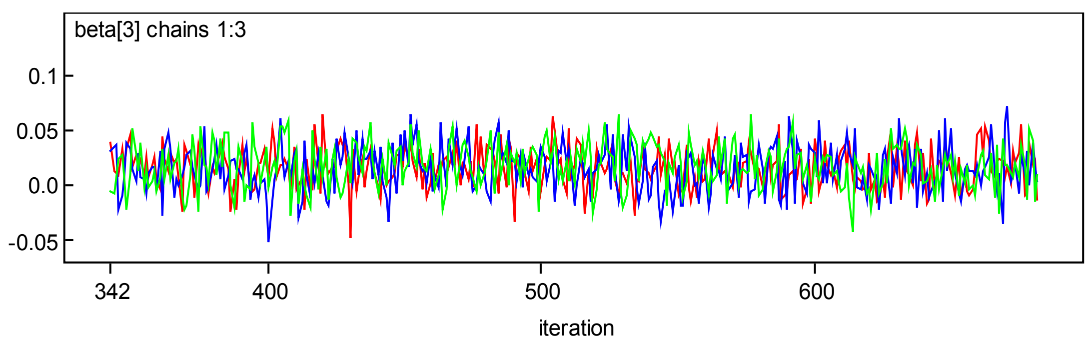

Figure 178. Trace Plot for Grade 3 Full Sample ELA Year 02

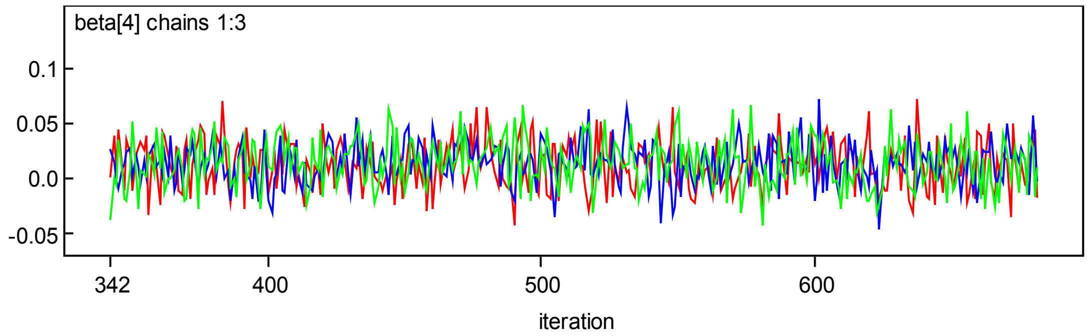

Figure 179. Trace Plot for Grade 3 Full Sample ELA Year 03

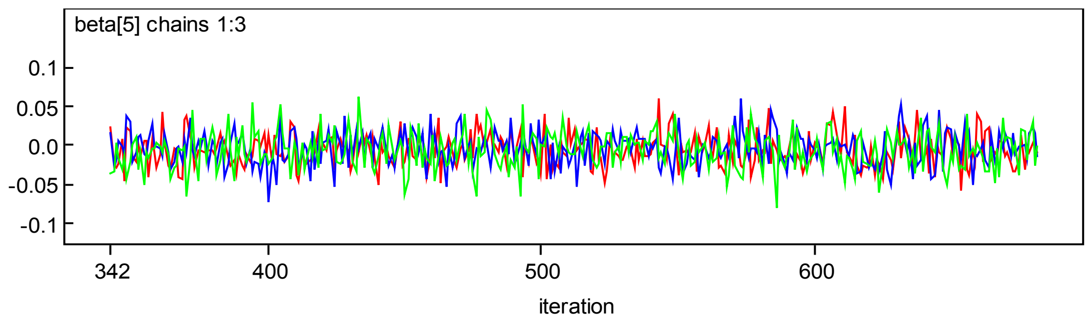

Figure 180. Trace Plot for Grade 3 Full Sample ELA Year 04

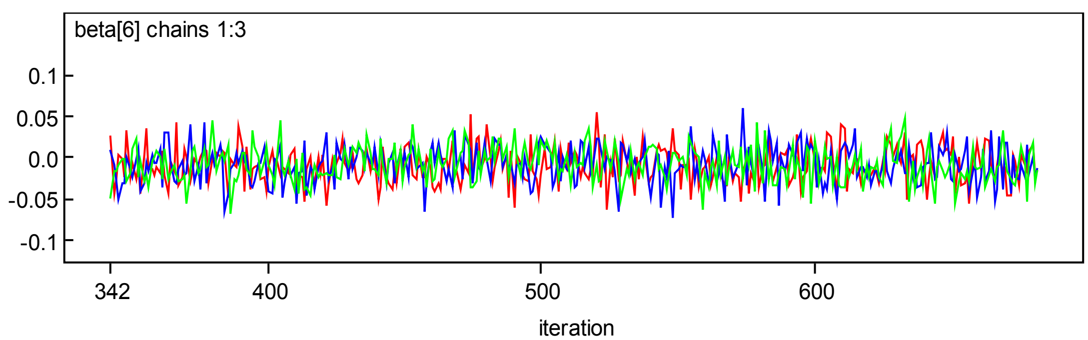

Figure 181. Trace Plot for Grade 3 Full Sample ELA Year 05

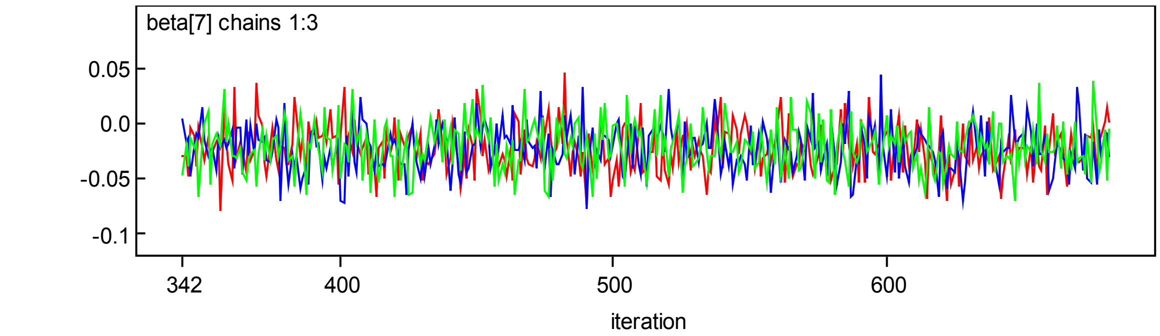

Figure 182. Trace Plot for Grade 3 Full Sample ELA Year 06

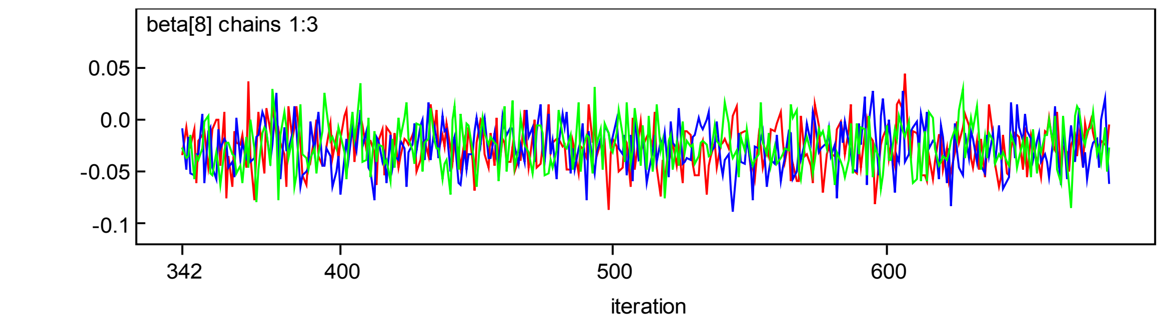

Figure 183. Trace Plot for Grade 3 Full Sample ELA Year 07

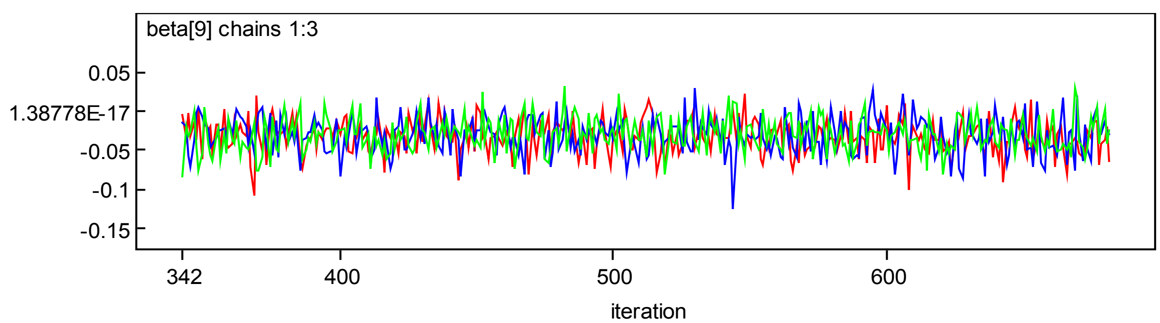

Figure 184. Trace Plot for Grade 3 Full Sample ELA Year 08

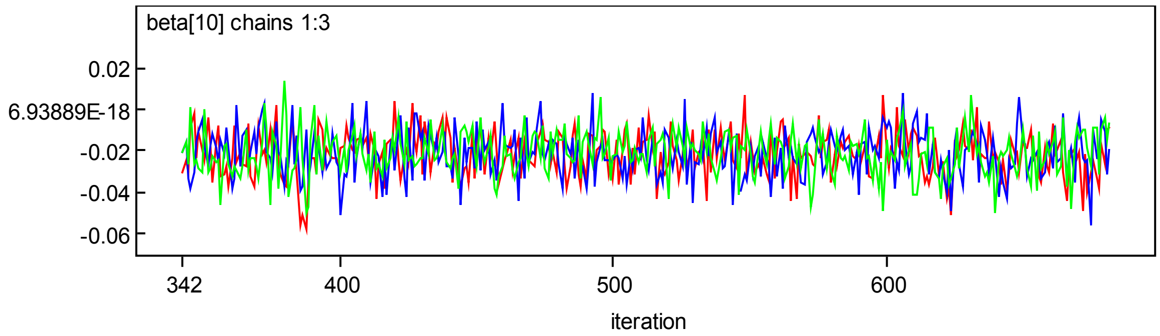

Figure 185. Trace Plot for Grade 3 Full Sample ELA Year 09

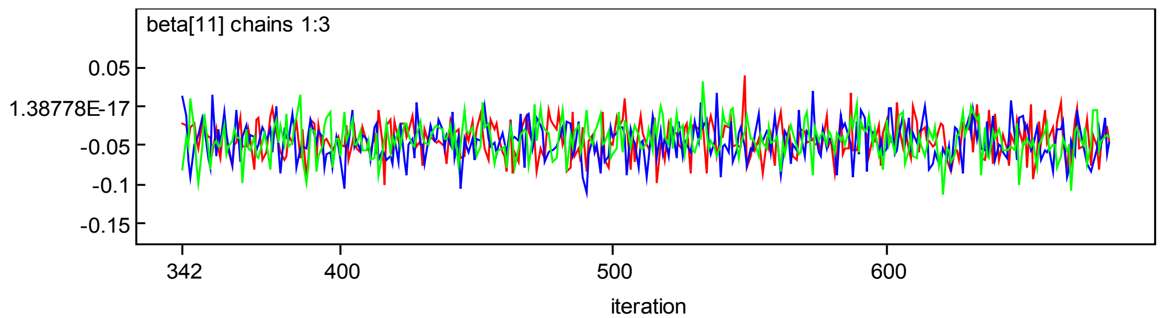

Figure 186. Trace Plot for Grade 3 Full Sample ELA Year 10

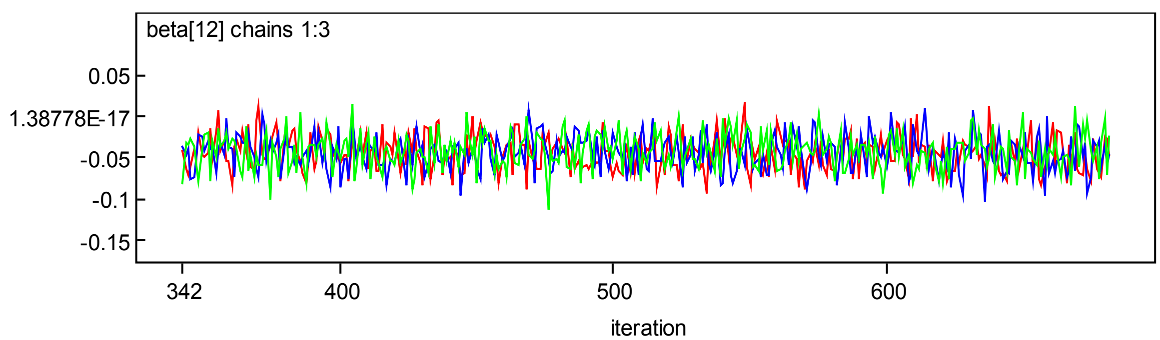

Figure 187. Trace Plot for Grade 3 Full Sample ELA Year 11

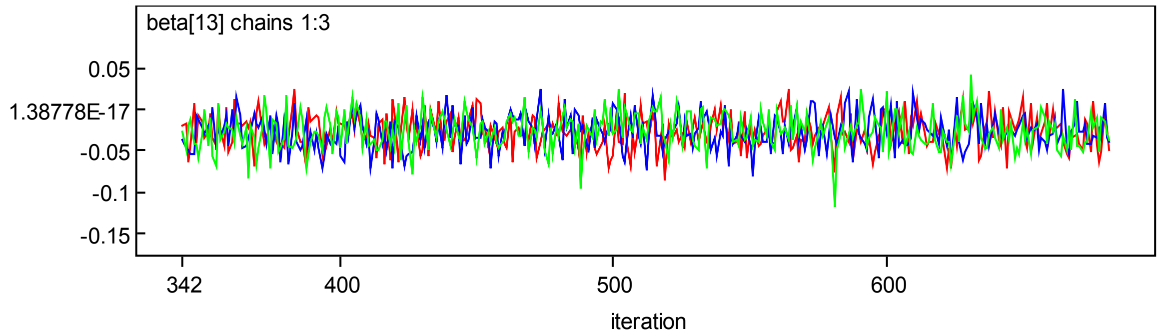

Figure 188. Trace Plot for Grade 3 Full Sample ELA Year 12

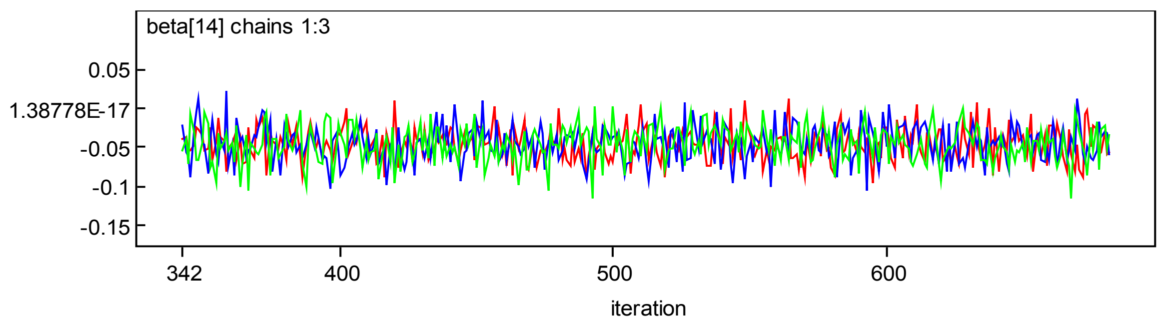

Figure 189. Trace Plot for Grade 3 Full Sample ELA Year 13

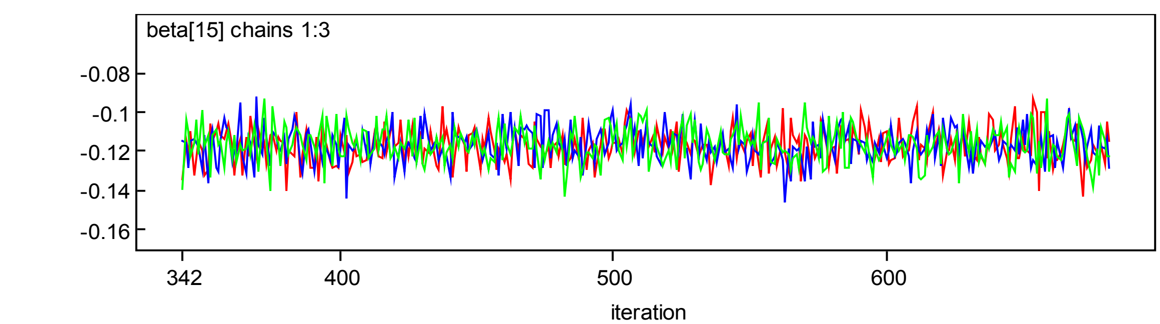

Figure 190. Trace Plot for Grade 3 Full Sample ELA Male

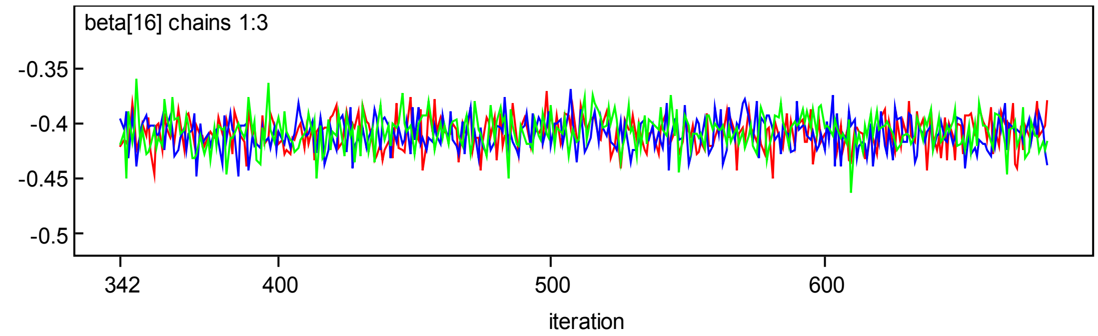

Figure 191. Trace Plot for Grade 3 Full Sample ELA Black

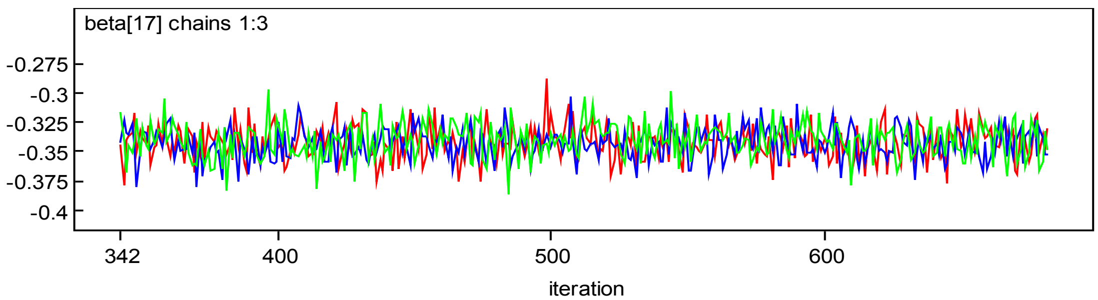

Figure 192. Trace Plot for Grade 3 Full Sample ELA Hispanic

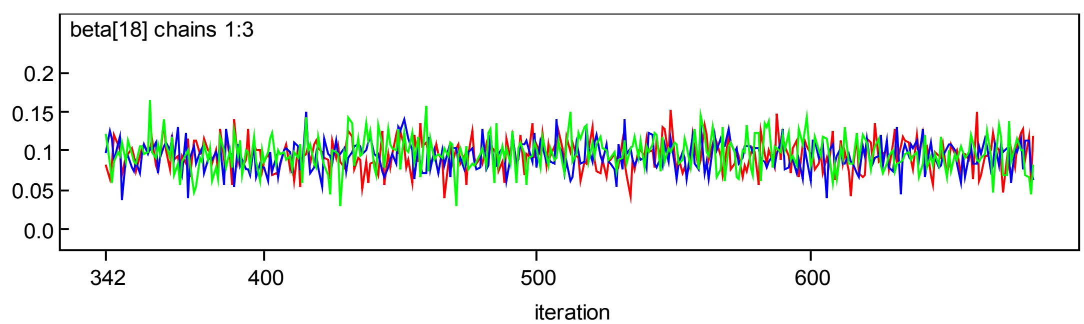

Figure 193. Trace Plot for Grade 3 Full Sample ELA Asian

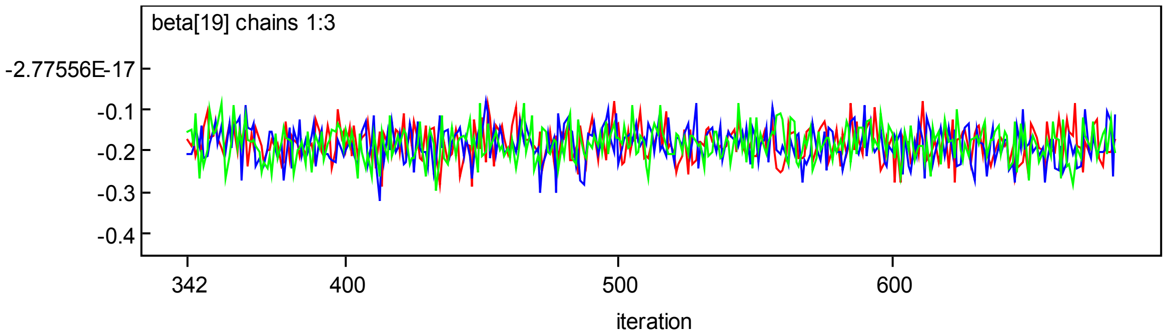

Figure 194. Trace Plot for Grade 3 Full Sample ELA Mixed

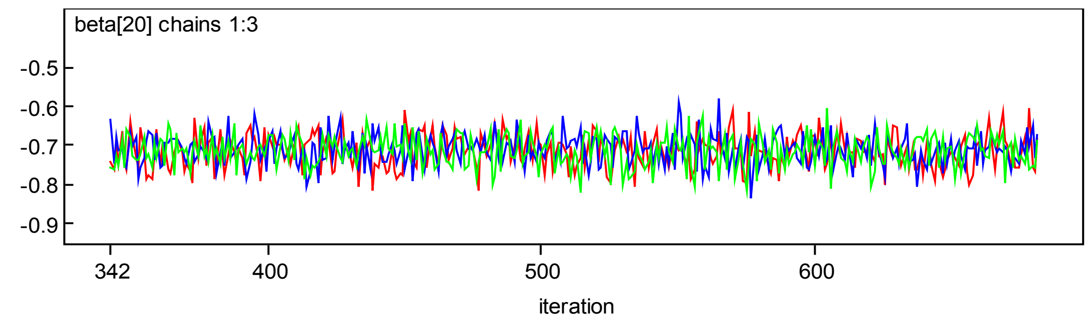

Figure 195. Trace Plot for Grade 3 Full Sample ELA Sped 1

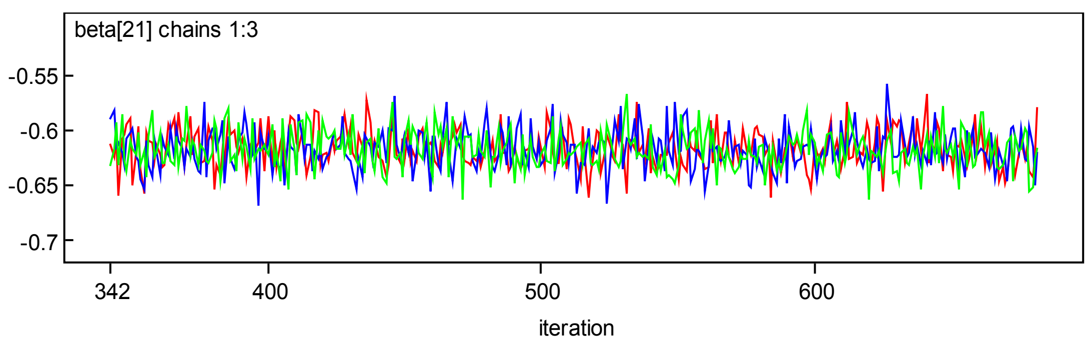

Figure 196. Trace Plot for Grade 3 Full Sample ELA Sped 2

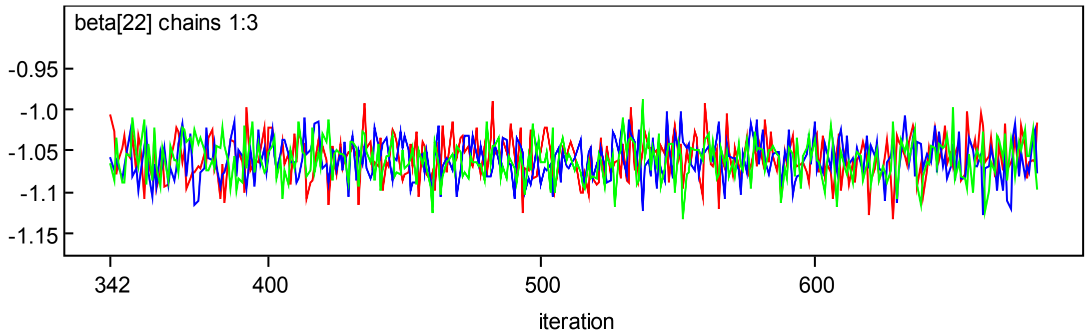

Figure 197. Trace Plot for Grade 3 Full Sample ELA Sped 3

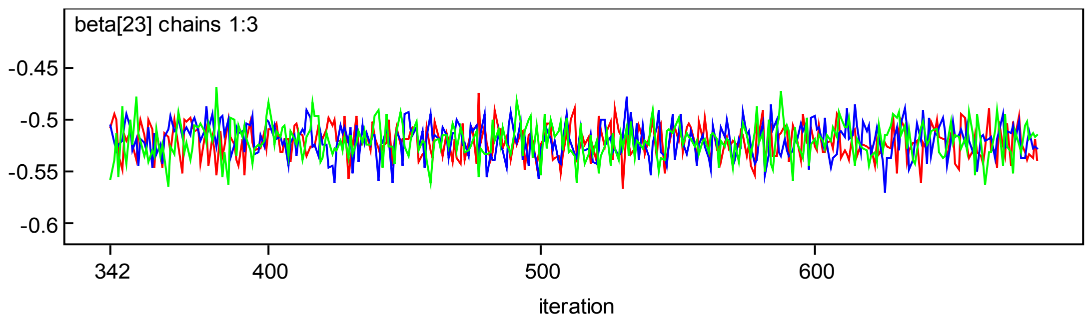

Figure 198. Trace Plot for Grade 3 Full Sample ELA Free Lunch

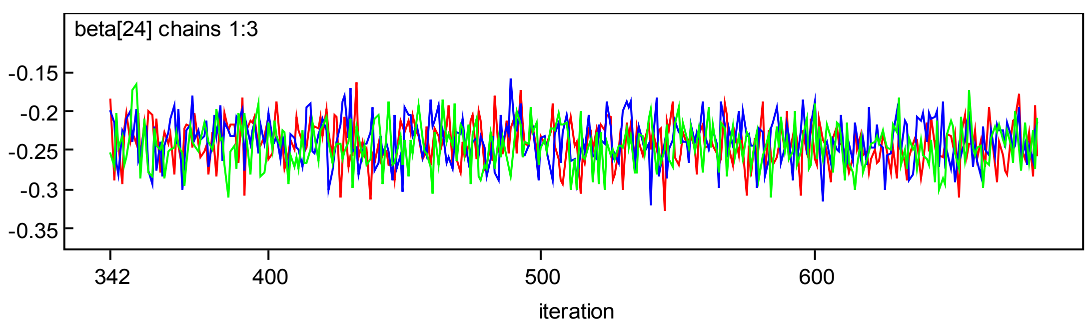

Figure 199. Trace Plot for Grade 3 Full Sample ELA Reduced Lunch

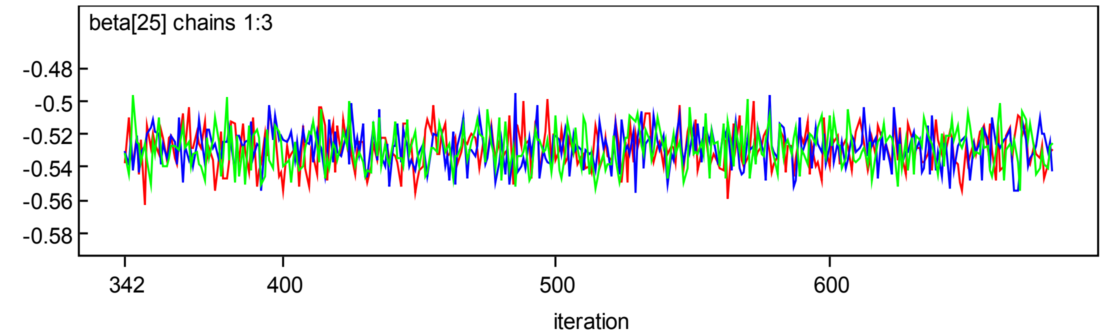

Figure 200. Trace Plot for Grade 3 Full Sample ELA Bilingual

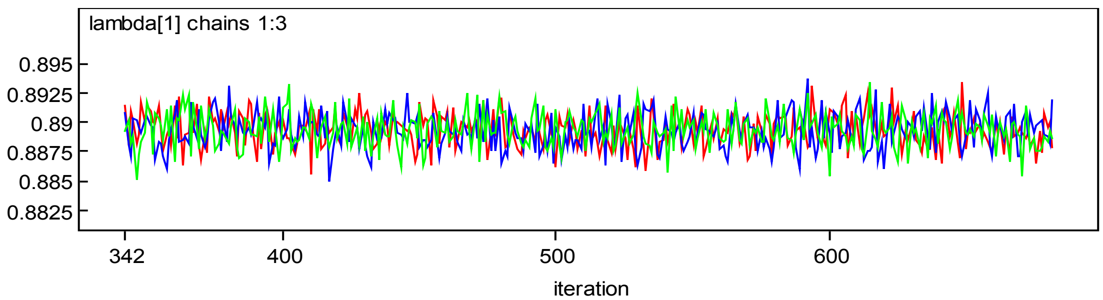

Figure 201. Trace Plot for Grade 3 Full Sample ELA Lambda 1

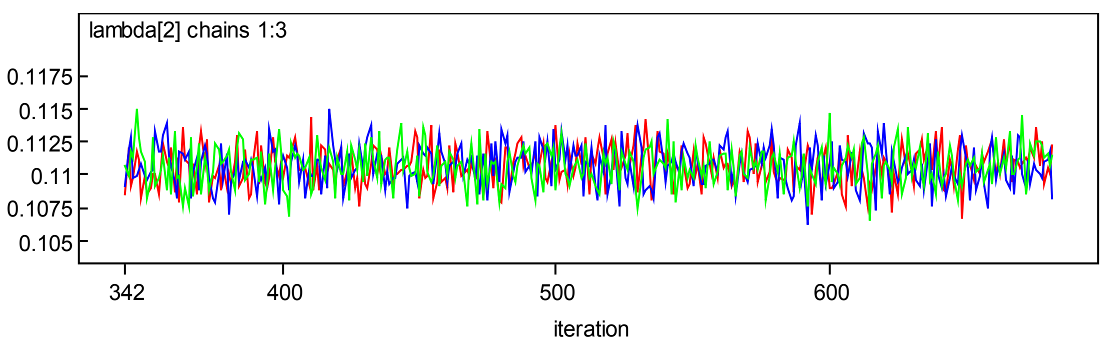

Figure 202. Trace Plot for Grade 3 Full Sample ELA Lambda 2

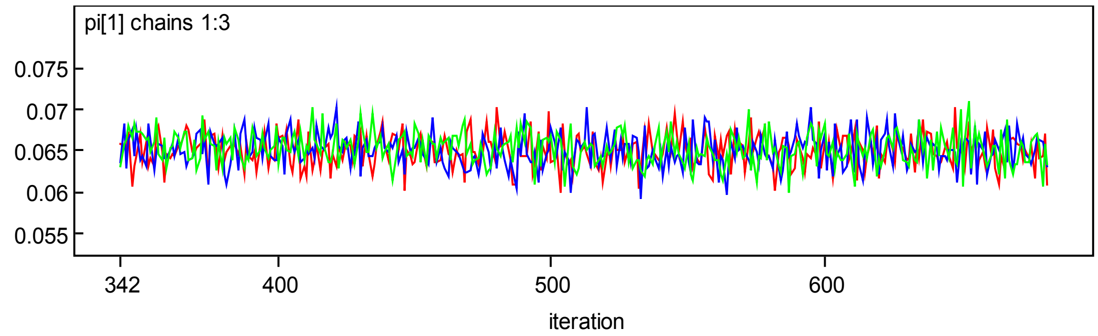

Figure 203. Trace Plot for Grade 3 Full Sample ELA Pi 1

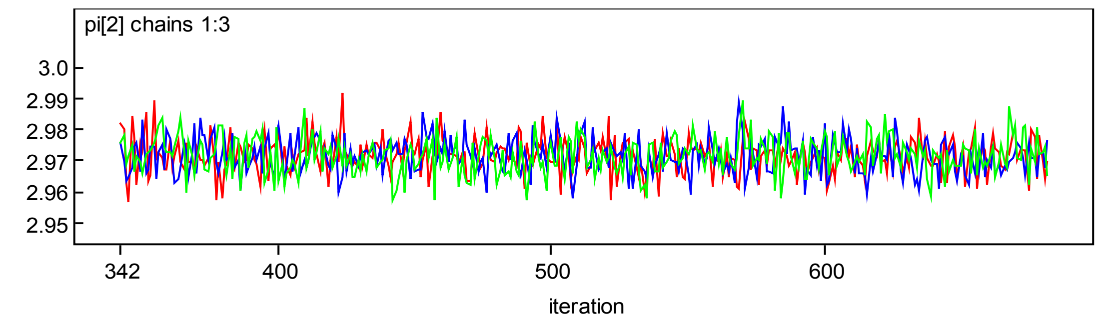

Figure 204. Trace Plot for Grade 3 Full Sample ELA Pi 2

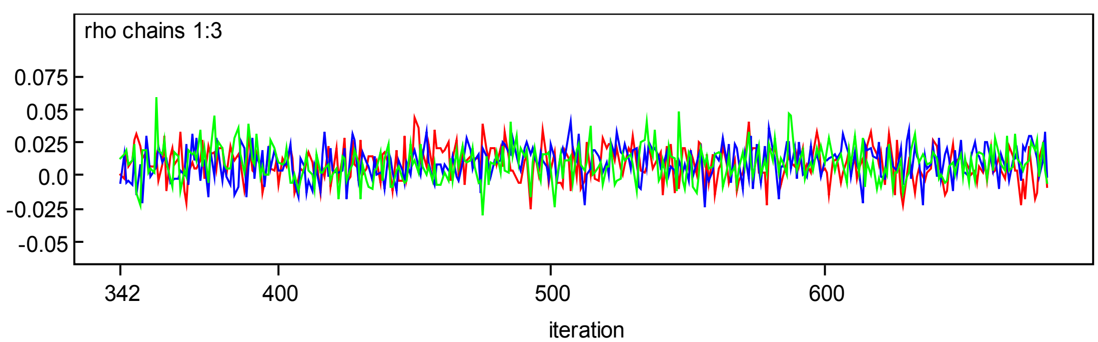

Figure 205. Trace Plot for Grade 3 Full Sample ELA Rho

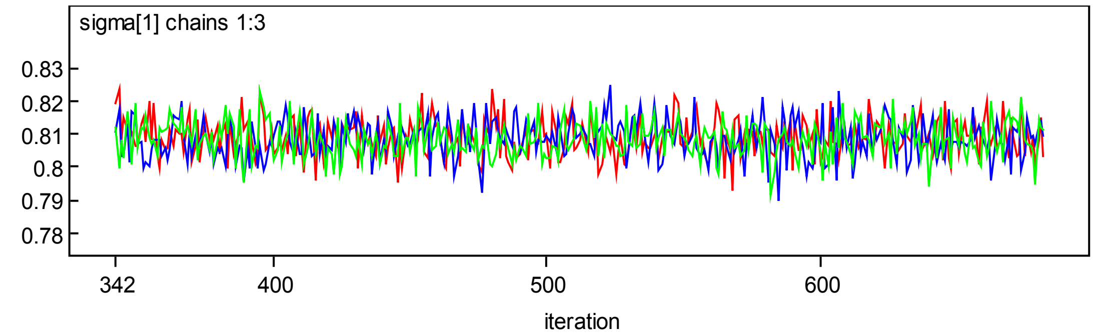

Figure 206. Trace Plot for Grade 3 Full Sample ELA Sigma 1

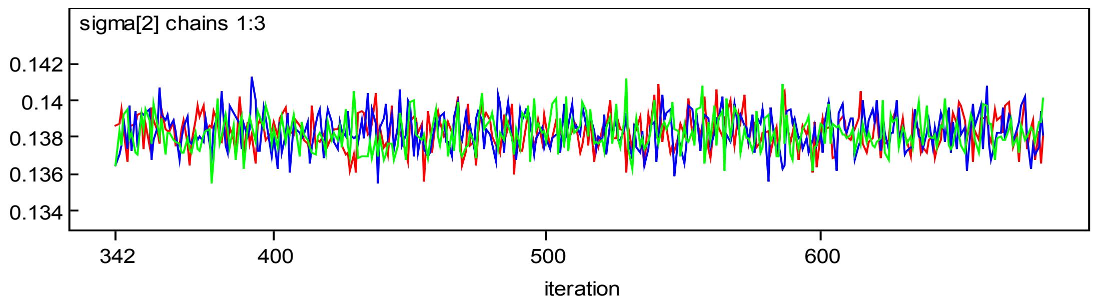

Figure 207. Trace Plot for Grade 3 Full Sample ELA Sigma 2

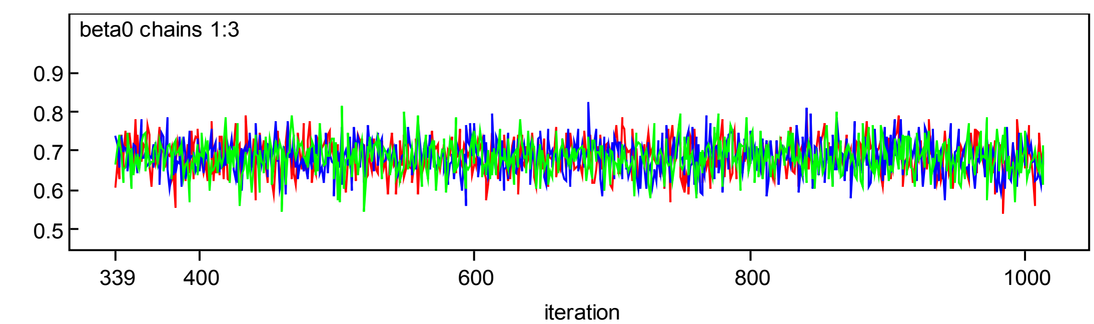

Figure 208. Trace Plot for Grade 3 Multilevel Full Sample Math Intercept

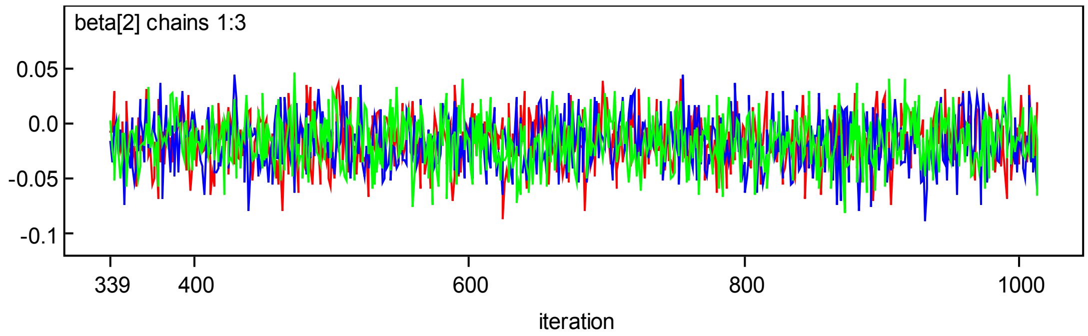

Figure 209. Trace Plot for Grade 3 Multilevel Full Sample Math Year 06

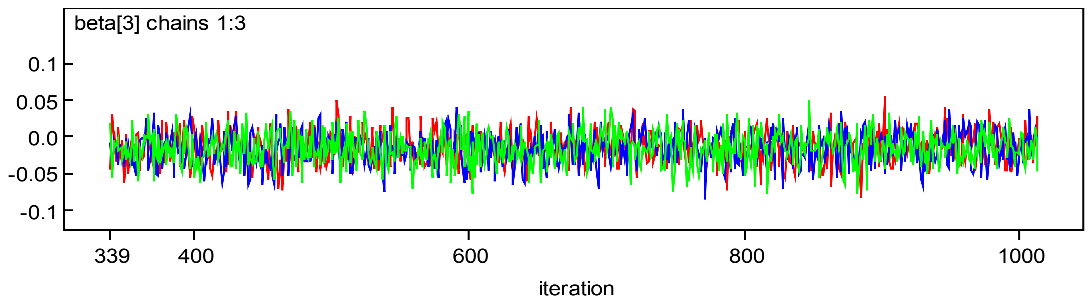

Figure 210. Trace Plot for Grade 3 Multilevel Full Sample Math Year 07

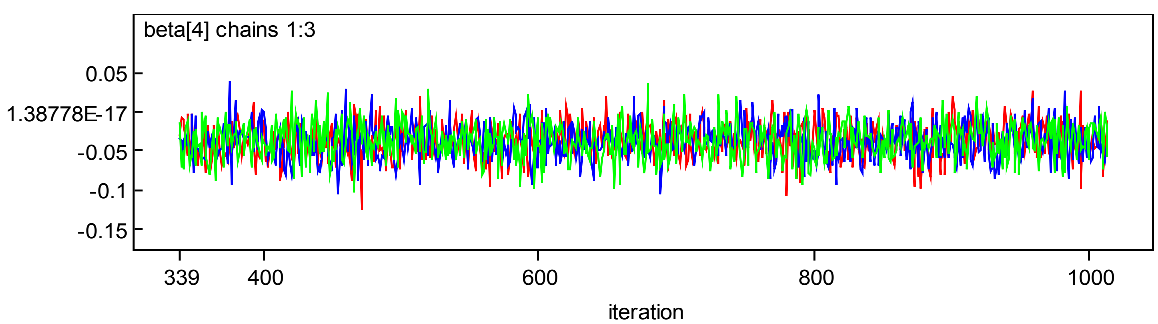

Figure 211. Trace Plot for Grade 3 Multilevel Full Sample Math Year 08

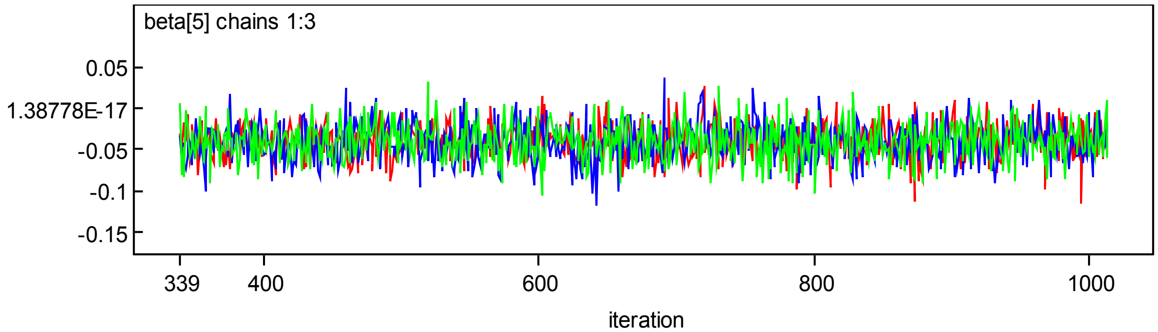

Figure 212. Trace Plot for Grade 3 Multilevel Full Sample Math Year 09

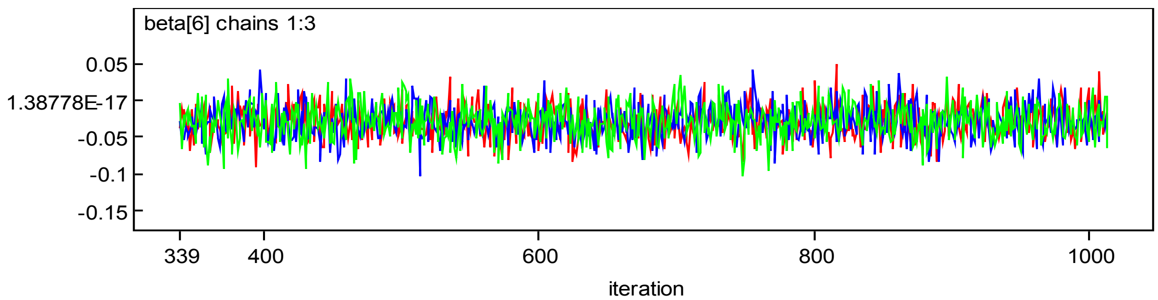

Figure 213. Trace Plot for Grade 3 Multilevel Full Sample Math Year 10

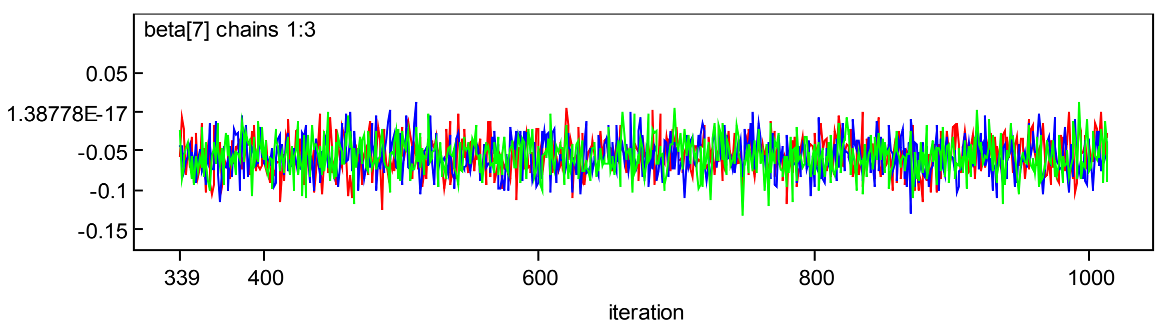

Figure 214. Trace Plot for Grade 3 Multilevel Full Sample Math Year 11

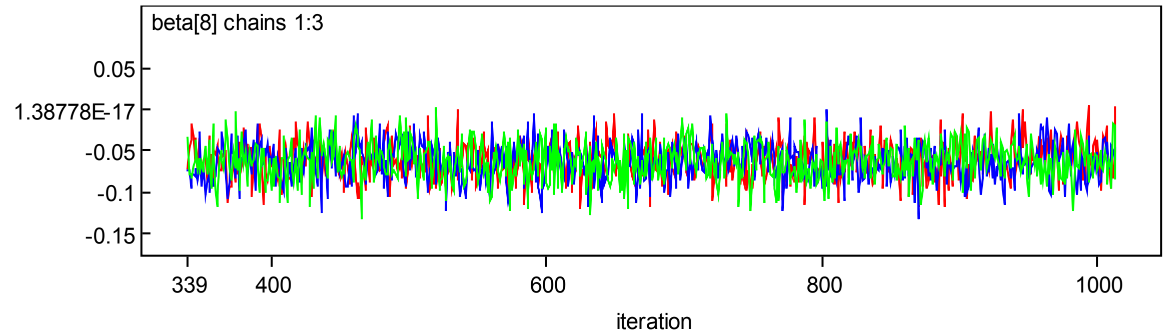

Figure 215. Trace Plot for Grade 3 Multilevel Full Sample Math Year 12

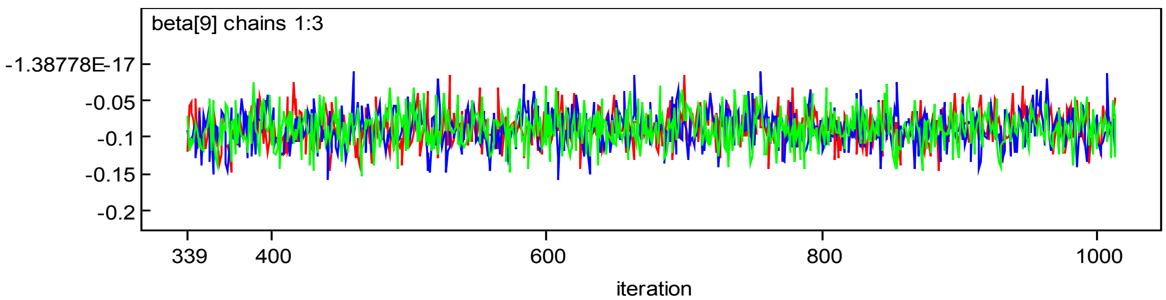

Figure 216. Trace Plot for Grade 3 Multilevel Full Sample Math Year 13

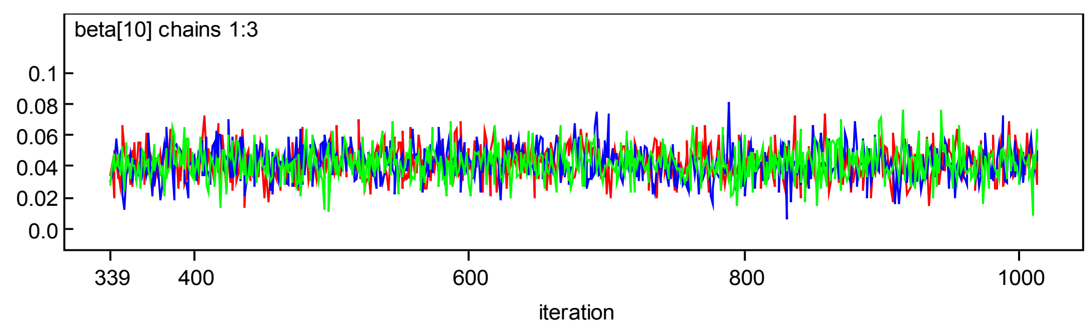

Figure 217. Trace Plot for Grade 3 Multilevel Full Sample Math Male

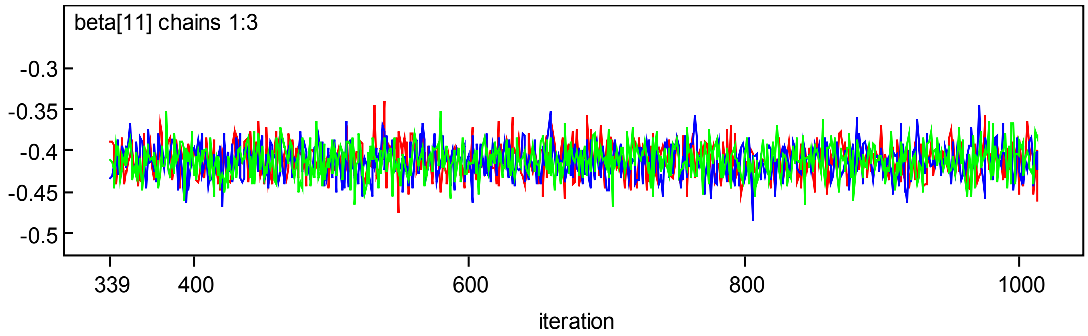

Figure 218. Trace Plot for Grade 3 Multilevel Full Sample Math Black

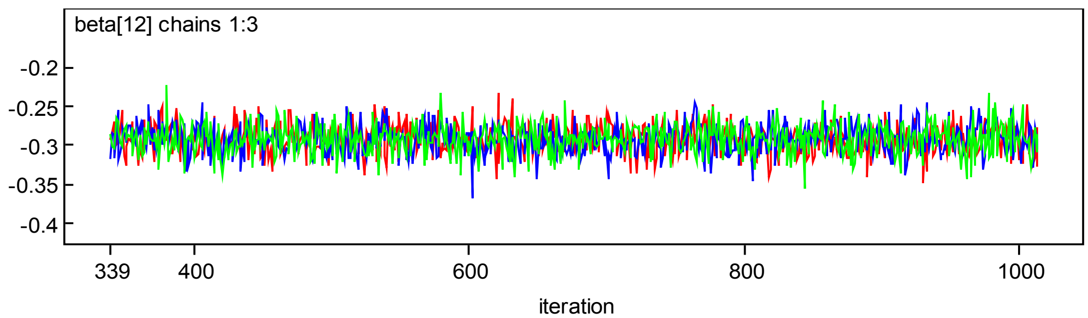

Figure 219. Trace Plot for Grade 3 Multilevel Full Sample Math Hispanic

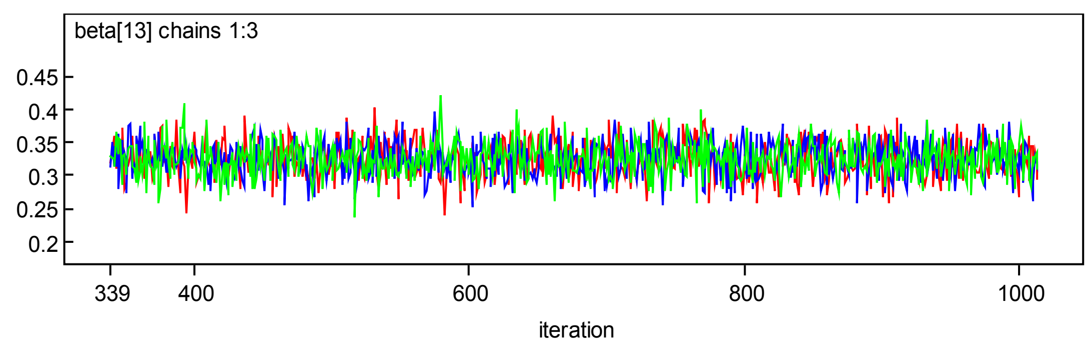

Figure 220. Trace Plot for Grade 3 Multilevel Full Sample Math Asian

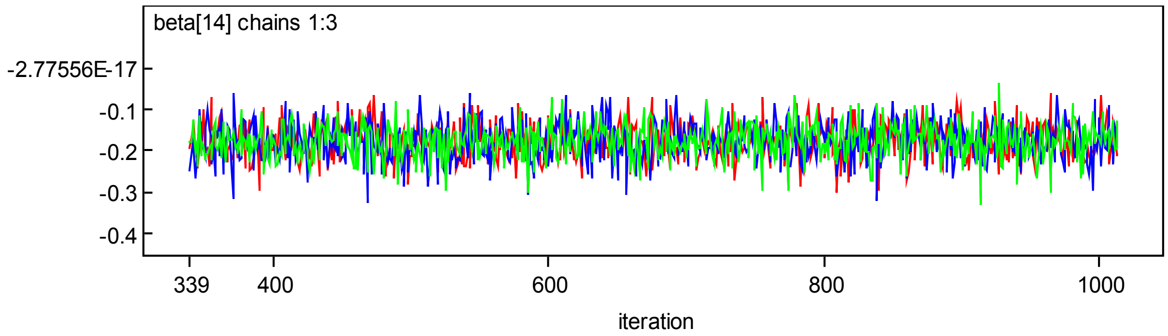

Figure 221. Trace Plot for Grade 3 Multilevel Full Sample Math Mixed

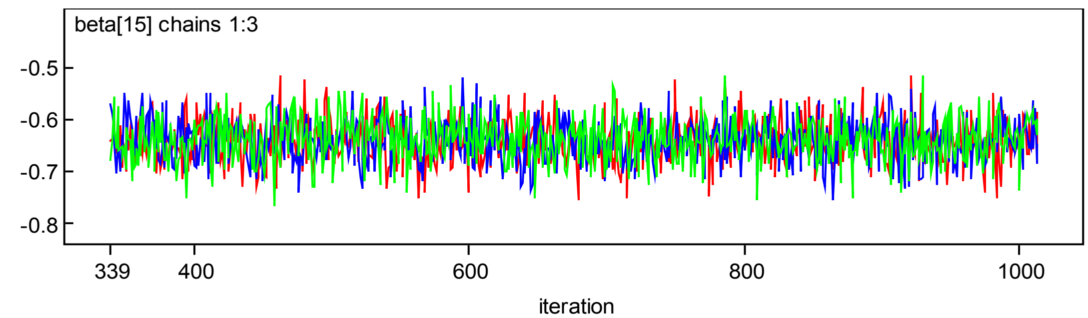

Figure 222. Trace Plot for Grade 3 Multilevel Full Sample Math Sped 1

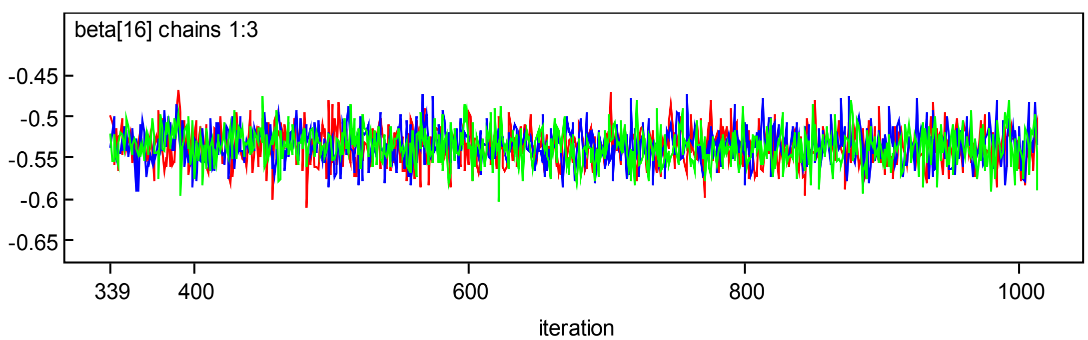

Figure 223. Trace Plot for Grade 3 Multilevel Full Sample Math Sped 2

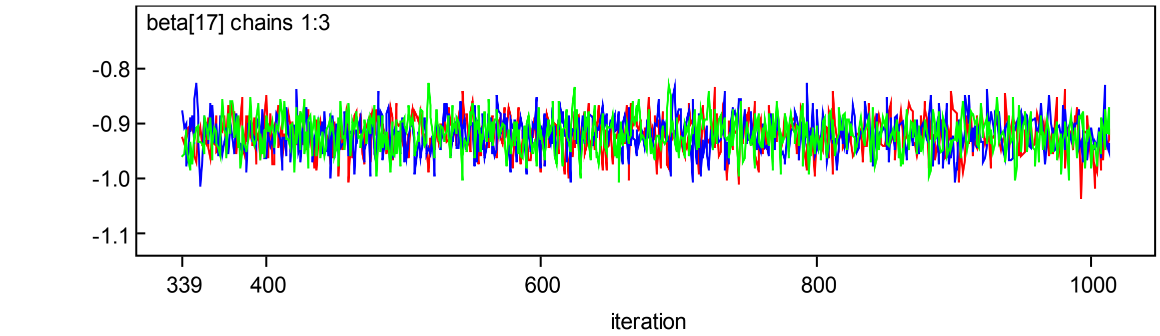

Figure 224. Trace Plot for Grade 3 Multilevel Full Sample Math Sped 3

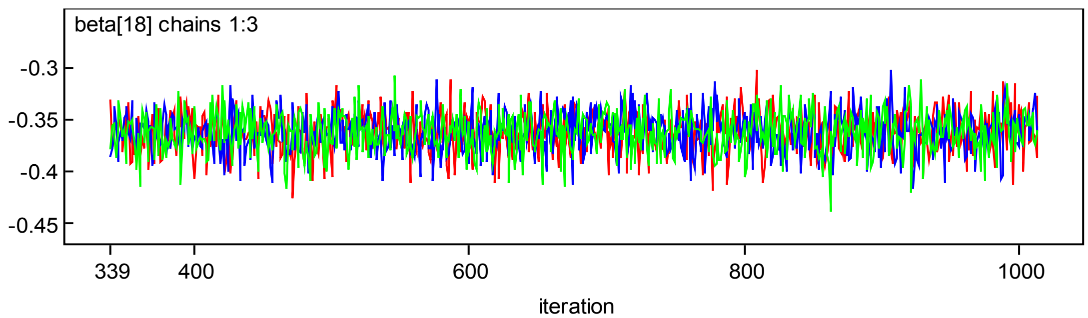

Figure 225. Trace Plot for Grade 3 Multilevel Full Sample Math Free Lunch

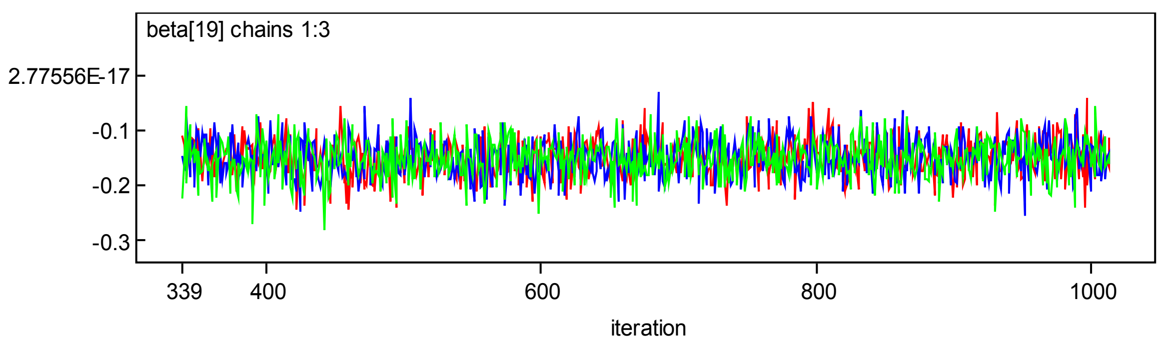

Figure 226. Trace Plot for Grade 3 Multilevel Full Sample Math Reduced Lunch

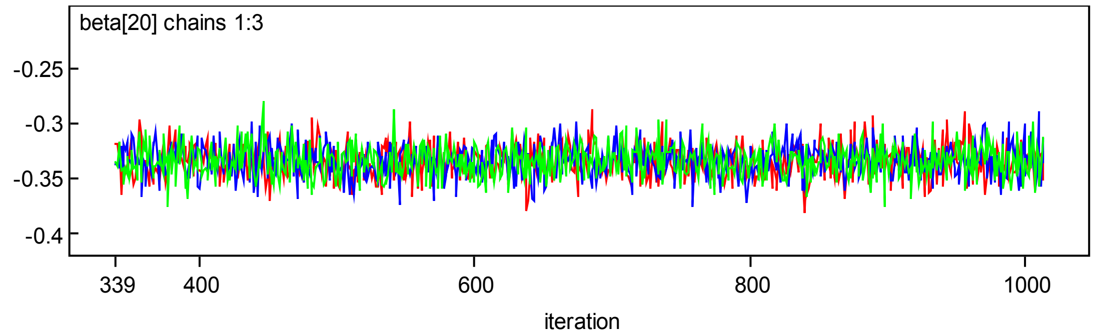

Figure 227. Trace Plot for Grade 3 Multilevel Full Sample Math Bilingual

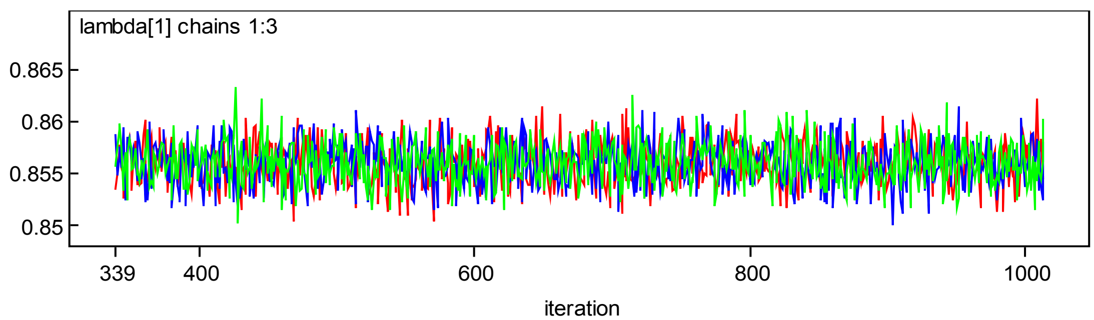

Figure 228. Trace Plot for Grade 3 Multilevel Full Sample Math Lambda 1

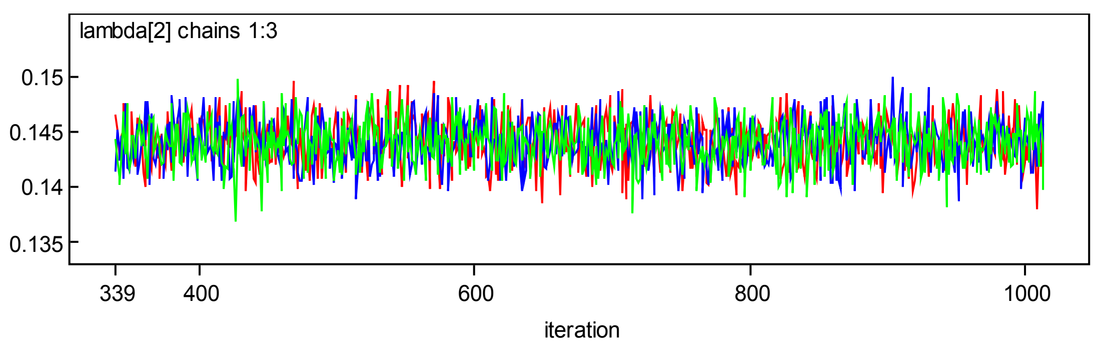

Figure 229. Trace Plot for Grade 3 Multilevel Full Sample Math Lambda 2

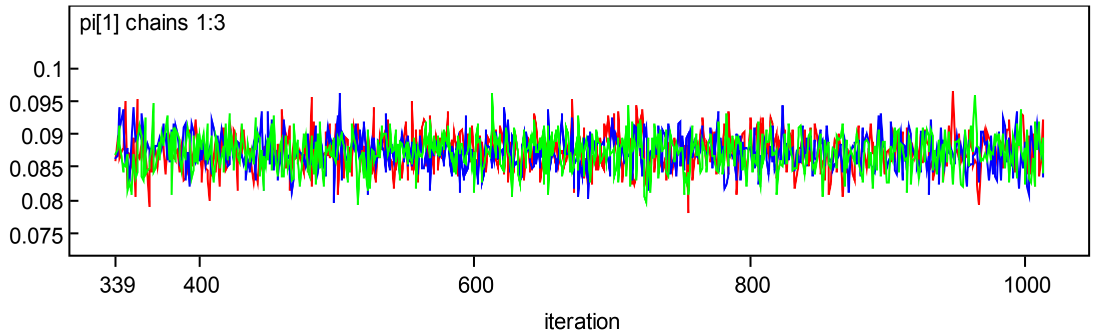

Figure 230. Trace Plot for Grade 3 Multilevel Full Sample Math Pi 1

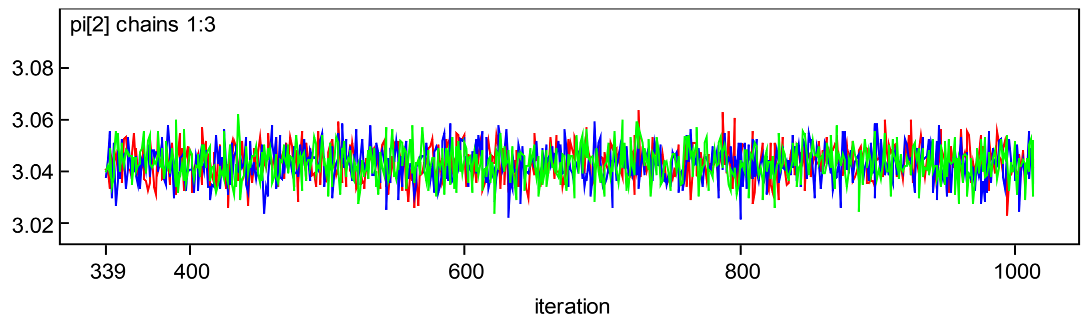

Figure 231. Trace Plot for Grade 3 Multilevel Full Sample Math Pi 2

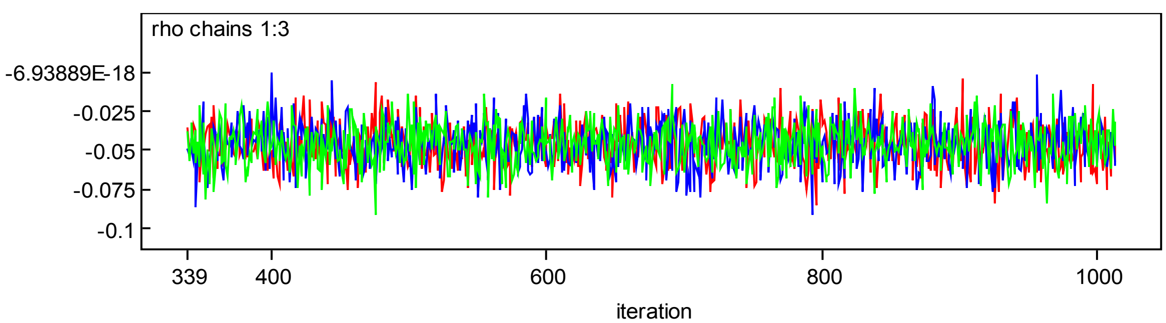

Figure 232. Trace Plot for Grade 3 Multilevel Full Sample Math Rho

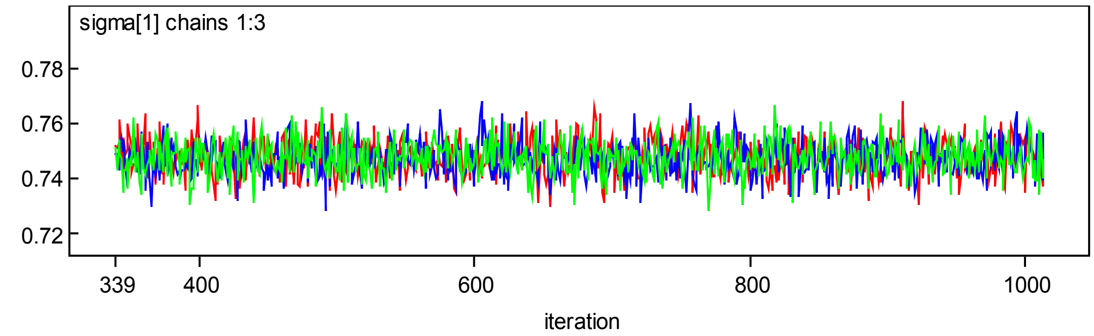

Figure 233. Trace Plot for Grade 3 Multilevel Full Sample Math Sigma 1

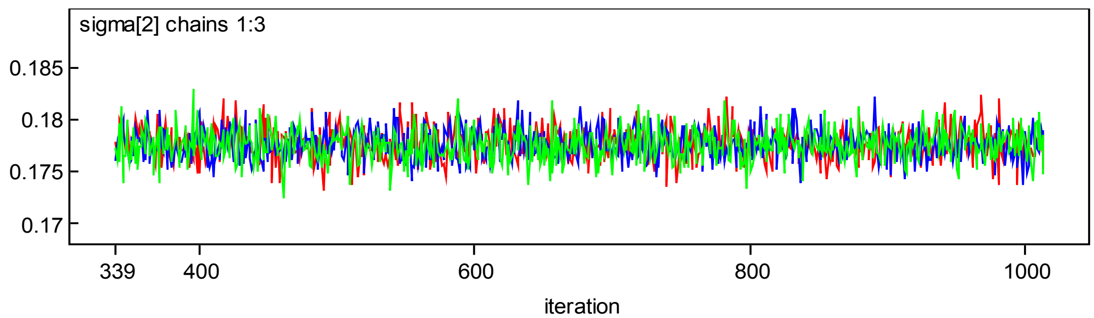

Figure 234. Trace Plot for Grade 3 Multilevel Full Sample Math Sigma 2