Uncertainty in Estimating Between-Teacher Variation for Value-Added Modeling: A Bayesian Perspective

Kellie Wills

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2013

Reading Committee:
Min Li, Chair
Robert Abbott
Margaret Plecki
John Miyamoto

Program Authorized to Offer Degree:
College of Education

University of Washington

Abstract

Uncertainty in Estimating Between-Teacher Variation for Value-Added Modeling:
A Bayesian Perspective

Kellie Wills

Chair of the Supervisory Committee:
Dr. Min Li
College of Education, Measurement, Statistics, and Research Design

The growing use of value-added modeling (VAM) in high-stakes personnel decisions implies that VAM teacher effect estimates can accurately differentiate higher- from lower-performing teachers. In fact, statistical power for these comparisons depends on the proportion of overall test score variation that is attributed to between-teacher differences – the intra-class correlation (ICC). This dissertation demonstrates innovative approaches to realistic treatment of ICCs as quantities estimated with uncertainty. Design priors, representing Bayesian prior beliefs about between-teacher ICC, were generated from thirty-one estimates from eight recently published VAM studies. The estimates from math test scores tended to be larger than estimates from reading test scores, and were also more variable, so separate priors were generated for math and reading estimates. This study introduces to the educational literature the use of fully Bayesian design priors to represent empirical evidence about ICC values. The fully Bayesian priors were

derived from two different distributional assumptions for the likelihood – Swiger's and Fisher's distributions – and compared to empirical Bayes (kernel density) priors. Analysis using simulated data sets supports the Fisher likelihood as a reasonable description of the distribution of the published estimates. Power analysis indicates that for either math or reading, empirically supported variation in the between-teacher ICC values has little effect on power to detect teachers who differ from average. However, somewhat better power can be expected for math teacher comparisons than for reading teacher comparisons. This study strengthens the evidence that the small contribution of between-teacher variance to total variance limits the utility of VAM for teacher comparisons. Acceptable (80%) power can be achieved for one year of data only for math scores, and only for a large threshold difference of a teacher from the average: more than a third of an average annual gain. The utility of Bayesian design priors for representing uncertainty about ICC extends to other prospective analyses for VAM, as well as to hierarchical analyses more generally.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Many thanks to my dissertation committee: Bob Abbott, John Miyamoto, Marge Plecki, and especially my advisor, Min Li. Min has been a model mentor, a constant source of intellectual stimulation and professional and emotional support. Special thanks to Dr. Rebecca Turner for her generosity in corresponding with a graduate student outside her field.

John Coltrane and Franz Joseph Haydn inspired me through long hours of writing; for me, the beauty of their music mirrors the beauty of mathematics. Throughout my life, my father and mother, James and Patricia Wills, have encouraged my academic endeavors. Words cannot express the importance of their support and my gratitude for it. The most fortunate aspect of my fortunate life has been my friendships. Thanks especially to Erik Anderson, Rita Chupalov, Jill Goldschneider, Andrew Jocuns, Thomas Salzman, and Suzie Scollon for their encouragement, humor, and wisdom throughout the Ph.D. process.

CHAPTER 1: Introduction

During the past 15 years, *value-added* modeling (VAM) approaches, which attempt to evaluate teacher, school, or program effectiveness through statistical modeling of student test scores, have increasingly influenced educational research and practice. As McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) wrote, "Enthusiasm for this approach stems in large part from the belief that it can remove the effects of factors not under the control of the school, such as prior performance and socioeconomic status, and thereby provides a more accurate indicator of school or teacher effectiveness than is possible when these factors are not controlled" (p. 68).

The goal of VAM is to estimate the effects on student test performance associated with particular school or teacher assignments, while controlling for prior test performance. Rubin, Stuart, and Zanutto (2004) discussed VAM as a *potential outcomes* (sometimes called *counterfactual*) approach that attempts to estimate the causal effect of one treatment (a teacher, for example) relative to others, when only one treatment can be observed per student. The term *value-added* highlights this counterfactual comparison of an observed teacher or school effect with the unobserved outcomes of other assignments. Statistically speaking, VAM is linear regression modeling with student test scores as outcome variables, and prior test scores, school and teacher assignments as predictors. Hierarchical linear modeling (HLM) (Raudenbush & Bryk, 2002) is commonly used to represent the dependence structure of educational data, in which students are clustered into classrooms that are grouped into schools.

As VAM has gained attention from researchers and policy makers, US states and districts have begun using VAM for school and teacher accountability purposes (National Research Council and National Academy of Education, 2010). These uses of VAM imply that VAM

teacher effect estimates can accurately differentiate higher- from lower-performing teachers. In fact, the accuracy of these comparisons depends on the proportion of overall test score variation that is attributed to between-teacher differences – that is, the intra-class correlation (ICC).

To date, the VAM literature has treated ICC values used in power calculations as fixed, known quantities. This dissertation will demonstrate innovative approaches to treating them more realistically, as quantities estimated with uncertainty. The analysis will be a hybrid of Bayesian and classical perspectives. The approach to parameter uncertainty will be Bayesian, using published estimates as prior information about the between-teacher ICC, and treating information about this parameter as a probability distribution for a random variable. The ultimate results will be expressed in terms of the traditional classical concept of power – in this case, power to detect teachers whose performance differs significantly from the average. However, these results will be Bayesian in the sense that power will be treated as a random variable with a probability distribution. This represents an innovation in the VAM literature because it enables meaningful statements about the *probability* of achieving power goals, given realistic assumptions about the structure of value-added modeling data. Specifically, the research questions asked in this dissertation will be as follows.

1. What does analysis of published estimates reveal about the proportion of variance in student test scores attributable to differences between teachers? In other words, what range and possible distributions of estimates for between-teacher ICC does the literature support?
2. Bayesian prior distributions (design priors) for the between-teacher ICC will be generated from the published estimates analyzed in question 1. What are the characteristics of design priors derived from:

a. Explicit distributional forms for the likelihood?

b. Empirical Bayes (Gaussian kernel) methods?

3. How does the form of the design prior affect the derived distribution for power to detect below- or above-average teachers?

CHAPTER 2: Review of the Literature

This chapter reviews prior scholarship in several areas pertinent to the current study's goals. First, an introduction to common value-added models and effect specifications will establish context for between-teacher ICC estimates in the VAM literature, treated in detail in Chapter 4. The review will then shift to the statistical literature for a Bayesian perspective on estimation uncertainty and its implications for research design. Finally, the chapter will describe strategies for incorporating estimation uncertainty into prospective analysis, using a hybrid of Bayesian and classical methodologies. Examples will be drawn from the medical statistics and educational research literatures.

Common Value-Added Models and Effect Specifications

This section will adopt the general mixed-effects value-added model of McCaffrey et al. (2004) as a framework for introducing VAM concepts. The outcome variable is a test score y_{ig} for student *i* in grade *g*. For a particular cohort, in the initial grade under consideration (grade 0), the model is as follows[1]:

$$y_{i0} = \mu_0 + \beta_0' x_i + \gamma_0' z_{i0} + \lambda_{i0}' \eta_0 + \phi_{i0}' \theta_0 + \varepsilon_{i0} \quad (1)$$

Here λ_{i0} and ϕ_{i0} are vectors of the proportions of time in grade 0 that student i spent in each of the district schools, and with each grade 0 teacher. Thus η_0 and θ_0 are vectors of school and teacher effects respectively. McCaffrey and colleagues specified these as normally distributed random effects, $\eta_0 \sim N\left(0, \sigma_{\eta_0}^2\right)$ and, $\theta_0 \sim N\left(0, \sigma_{\theta_0}^2\right)$, but they may also be specified as fixed effects, as described in the next section. The model may include student- or classroom-level covariates, both time invariant (x_i), such as race, and time-varying (z_{i0}), like test accommodation status. The errors are assumed $\varepsilon_0 \sim N\left(0, \sigma_{\varepsilon_0}^2\right)$.

[1] McCaffrey et al.'s equations have been corrected for apparent typographical errors in the paper.

The models for future years of data include effects of prior year teachers and schools on the current year's test score:

$$y_{i1} = \mu_1 + \beta_1' x_i + \gamma_1' z_{i1} + (\omega_{10}\lambda_{i0}'\eta_0 + \lambda_{i1}'\eta_1) \quad (2)$$
$$+ (\alpha_{10}\phi_{i0}'\theta_0 + \phi_{i1}'\theta_1) + \varepsilon_{i1}$$
$$y_{i2} = \mu_2 + \beta_2' x_i + \gamma_2' z_{i2} + (\omega_{20}\lambda_{i0}'\eta_0 + \omega_{21}\lambda_{i1}'\eta_1 + \lambda_{i2}'\eta_2)$$
$$+(\alpha_{10}\phi_{i0}'\theta_0 + \alpha_{21}\phi_{i1}'\theta_1 + \phi_{i2}'\theta_2) + \varepsilon_{i2}$$

and so on. The α/ω parameters represent the contribution of prior grade teachers/schools to the current grade's test score; other parameters are as described in the previous paragraph.

The general model of equations 1-3 provides a framework for understanding several commonly used models that can be thought of as special cases (McCaffrey et al., 2004; Rothstein, 2010). Variants of the general model with additional restrictions on the parameters can be estimated as standard hierarchical linear models (HLMs) with students nested within teachers (Raudenbush & Bryk, 2002). One example is a gain score HLM, in which the dependent variable is the difference between a student's most recent test score and a previous score. Schochet and Chiang (2013) used a four-level gain score HLM for their work on error rates in VAM, to be described later in the chapter. The dependent variable g_{itjk} is the difference between the current year and previous year test scores.

Level 1: Students i = 1, …, n (3)

$$g_{itjk} = \xi_{jk} + \varepsilon_{itjk}, \varepsilon_{itjk} \sim N(0, \sigma_\varepsilon^2)$$

Level 2: Classrooms t = 1, …, c

$$\xi_{tjk} = \tau_{jk} + \omega_{tjk}, \omega_{tjk} \sim N(0, \sigma_\omega^2)$$

Level 3: Teachers j = 1, …, m

$$\tau_{jk} = \eta_k + \theta_{jk},\ \theta_{jk} \sim N\left(0, \sigma_\theta^2\right)$$

Level 4: Schools k = 1, …, s

$$\eta_k = \delta + \psi_k,\ \psi_k \sim N\left(0, \sigma_\psi^2\right)$$

Schochet and Chiang's VAM involves several additional assumptions beyond those of the general model in equations 1 and 2. The gain score as dependent variable assumes "complete persistence" of teacher and school effects into subsequent grades (all α and ω terms in the general model are assumed to be 1). This implies that prior teachers and schools have no impact on this year's score *gain*, because the terms from prior teachers/schools cancel out in the subtraction. Schochet and Chiang also omit student-level covariates, under the common assumption that "a student's history of test performance substitutes for omitted background variables" (Ballou, Sanders, & Wright, 2004, p. 37).

VAM school and teacher effects may be specified either as normally distributed random effects or as fixed effects – that is, indicator or "dummy" variables for individual teachers or schools. The choice of fixed or random effects has generally been split along disciplinary lines: researchers trained as statisticians generally prefer random effects (e.g., McCaffrey et al., 2004; Nye, Konstantopoulos, & Hedges, 2004), while those trained as econometricians tend to prefer fixed effects (e.g., Jacob & Lefgren, 2008; Kane, Rockoff, & Staiger, 2008; Kane & Staiger, 2008; Koedel & Betts, 2011; Rockoff, 2004; Rothstein, 2010). As Lockwood and McCaffrey (2007) describe the situation, "Common approaches in the statistical literature focus on modeling… unobserved heterogeneity as part of the error structure for the data using random effects or mixed models… Economists have tended to focus more [on] the potential biasing effects of unobserved heterogeneity and fixed effects approaches to remove that bias under the appropriate assumptions. Fixed effects approaches introduce parameters for each individual as part of the model mean structure, rather than the error structure" (p. 225). The "potential biasing

effects" mentioned by Lockwood and McCaffrey may result from nonrandom assignment of students to teachers and of students and teachers to schools, to be discussed in Chapter 4.

Estimates of Between-Teacher Variation in the VAM Literature

Use of VAM in accountability contexts raises a fundamental and controversial issue: how much of the variability in test scores can be attributed to differences between teachers? This question is of interest in its own right, since it contributes to our understanding of the role of teachers in student learning.[2] The current work, however, focuses on a specific policy-related implication of the question: between-teacher variation affects the feasibility of statistically ranking or classifying teachers. These statistical teacher comparisons have become critical in the accountability contexts in which VAM is being used (National Research Council and National Academy of Education, 2010).

Intense research interest in VAM has produced an array of between-teacher variance estimates from various U.S. geographical areas and a complex variety of value-added model specifications. In their recent review of estimates, Hanushek and Rivkin (2010) discussed estimates from ten published and unpublished studies. The average of their reported variance estimates for math is 0.023, and for reading 0.012, standardized to unit total variance of test scores. Hanushek and Rivkin asserted, "The magnitudes of these estimates support the belief that teacher quality is an important determinant of school quality and achievement. For example, the math results imply that having a teacher at the twenty-fifth percentile as compared to the seventy-fifth percentile of the quality distribution would mean a difference in learning gains of roughly 0.2 standard deviations in a single year… The magnitude of such an effect is large both relative to typical measures of black-white or income achievement gaps of 0.7-1 standard

[2] VAM proponents conceptualize between-teacher variance as variation in teacher "effectiveness" or "quality", implying that teachers fully control performance differences. However, this variation could also include aspects of the instructional context that teachers cannot control, such as access to materials or availability of support personnel.

deviation and compared to methodologically compelling estimates of the effects of a ten student reduction in class size of 0.1-0.3 standard deviations" (p. 268).

Since publication of Hanushek and Rivkin (2010), additional estimates from new studies have become available. Also, although these authors raised insightful methodological concerns, they did not address the potential impact of different modeling strategies on the estimates. Chapter 4 will provide an updated review of estimates in the literature, with a focus on model specification. The next section will discuss methods of assessing the implications of between-teacher variation for the accountability goals of VAM.

Prospective Analysis for VAM

Schochet and Chiang (2013) observed that a typical use of VAM for accountability "classifies each teacher into a performance category based on the *t* statistic from testing the null hypothesis that the teacher's performance is equal to the average performance in a reference group" (p. 7). Meaningful classification requires acceptable statistical power, that is, a reasonable probability that the null hypothesis will in fact be rejected for a teacher who differs from average. Power in turn depends on the proportion of the overall variation in test scores that can be attributed to differences between teachers – that is, the intra-class correlation (ICC).

The ICC has two common interpretations:

1. The ICC can be thought of as the correlation of outcomes for students of the same teacher.
2. The ICC can also be interpreted as the proportion of the total variation in test scores explained by differences between teachers. The current work will use a three-level hierarchical model, with students nested within teachers nested within schools. For this type of model, the between-teacher ICC is calculated as:

$$\rho_T = \frac{\sigma^2_{teacher}}{\sigma^2_{teacher}+\sigma^2_{school}+\sigma^2_{residual}} \quad (4)$$

The power of the *t*-test described above depends on the ICC: the larger the between-teacher variation relative to the other sources of variation, the more likely that a given difference of a teacher from average can be detected.

An important tool for research design is *prospective* analysis, conducted prior to collecting data to evaluate whether a design will achieve study goals (such as reasonable power). Prospective analysis takes into account prior beliefs about the parameters driving the data, as well as the structure and sample sizes of the design (Kruschke, 2011). An important recent prospective analysis for VAM is Schochet and Chiang (2013), which predicts error rates for the four-level HLM described in section 2.1. From the normality assumptions of classical hypothesis testing, Schochet and Chiang derived the following expression for Type II error rate for testing the null hypothesis that the teacher's performance is equal to the average performance in a reference group:

$$1 - \beta = 1 - \Phi\left[\frac{|T|\sigma}{2\sqrt{V_{est}}}\right] \quad (5)$$

In equation 5, β is the power (so that 1 - β is the Type II error rate), σ is the total variance of scores, and V_{est} is the variance of the teacher effect estimator, which will be described in Chapter 3. Schochet and Chiang set the threshold T to represent an educationally meaningful difference of a teacher from the average. They first expressed average annual growth per grade in units of standard deviations of gain scores. For reading, they used an average annual growth estimate of 0.65 standard deviations, while for math they used 0.94 standard deviations. Based on these values, they chose threshold values of 0.1, 0.2, and 0.3 standard deviations. "A .2 value represents 31% of an average annual gain score in reading, or about 4 months of reading growth attained by a typical upper elementary student; in math, it represents 21% of an average annual

gain score, or about 3 months of student learning" (p. 157).[3] In other words, they assumed that an above-average teacher achieves greater gains in the same instructional time.

Schochet and Chiang derived variance values for the analysis from point estimates of ICC. For the between-classroom, between-teacher, and between-school ICCs, their point estimates were the means of estimates from 15 recent published and unpublished studies. Between-classroom variance was estimable from studies that used data sets including multiple years (multiple classrooms) taught by the same teacher; for other studies, between-classroom variance was imputed (Schochet & Chiang, 2010).

Schochet and Chiang's analysis indicates that even for the largest threshold (0.3 standard deviations), at least three years of gain scores for a teacher (three classrooms, including prior year test scores for each student) are required to attain power of at least 80%, often considered minimum acceptable power. For a threshold of 0.2 standard deviations, 5 years of data are required, while for 0.1 standard deviations, not even 10 years are sufficient (Schochet & Chiang, 2013).

Schochet and Chiang treated their ICCs as fixed, known quantities. However, Chapter 4 will demonstrate that published estimates of between-teacher ICC show considerable variation. We face a great deal of uncertainty about the actual ICC values appropriate for use in a prospective analysis. This raises the question of the impact of different possible values of ICCs on power. Schochet and Chiang performed some analysis of the sensitivity of their results to different ICC values; however, this involved simultaneous manipulation of the between-teacher and between-classroom ICCs, so that the impact of changes in the between-teacher ICC alone

[3] For reading, 0.2/0.65 = 0.31; for math, 0.2/0.94 = 0.21. Recall that according to Hanushek and Rivkin (2010), 0.2 standard deviations is approximately the difference between the 25th and 75th percentiles of teachers. According to Schochet and Chiang (2013), 0.2 standard deviations is the difference between the 50th and 82nd percentiles of teachers.

was obscured. Also, the alternative values selected for analysis were not chosen to represent the range of values supported by the literature. An opportunity remains for further VAM prospective analysis that takes into account our limited knowledge about the between-teacher ICC and the literature support for a range of possible values of this parameter.

Bayesian Prospective Analysis

This dissertation study will use Bayesian analysis as a strategy for incorporating uncertainty about between-teacher ICC into prospective analysis. Bayesian methods have become a well-established approach to parameter uncertainty, because they treat information about model parameters as probability distributions (Gelman & Hill, 2007). These distributions provide a much fuller description of more and less likely parameter values than point estimates or confidence intervals. The following discussion of Bayesian prospective analysis will use the framework and notation of DeSantis (2007), who called it *prior predictive* analysis.

Bayesian estimation calculates or simulates a posterior distribution for model parameters θ, given the data y, from a prior distribution and the sampling distribution of the data:

$$p(\theta|y) \propto p(y|\theta)\pi(\theta) \tag{6}$$

The posterior distribution is proportional to a combination of:

1. The sampling distribution or likelihood of the data, $p(y|\theta)$. This is the probability distribution we expect our observations to follow, dependent on parameters θ.
2. A prior distribution of the parameters, $\pi(\theta)$. This represents our beliefs about θ, expressed as a probability distribution.

As a simple example, consider Bayesian prospective analysis for a political poll, in which there is a choice between two candidates (Kruschke, 2011). Voters are to be polled by simple random sampling, with the goal of estimating the percentage of voters supporting one candidate

(p). In this example, the sampling distribution is Binomial(p, n), with n the number of voters polled. In cases where little is known about the parameters, the prior $\pi(\theta)$ may have a "non-informative" distribution in which all possible values have the same probability of occurrence. In this example, an appropriate choice would be Uniform[0, 1].

A fully Bayesian prospective analysis would be performed by simulating data as follows:

1. Draw a random value of the parameter from the prior. In the poll example, we might draw one value of p from a prior incorporating specific knowledge of the situation (such as the results of previous polls).
2. Given the parameter value from step 1, draw a random sample of data from the sampling distribution. In the poll example, we would draw a Binomial sample of size n with parameter p.
3. For the data from 2, generate a Bayesian posterior distribution for p, representing the information this particular data set provides about p.
4. Repeat steps 1-3 many times to generate a large number of data sets and posterior distributions for p.

The many data sets generated in step 3 are in effect many simulated repetitions of the same research study, and the posterior distributions in step 4 represent estimates calculated from the data sets. We then consider the proportion of simulated data sets in which a research goal is achieved. One such research goal might be excluding a particular parameter value. In the poll example, suppose our concern is whether $p > 0.5$, and our prior belief indicates a distribution for p centered around, say, 0.65. The posterior distributions generated in step 3 can be summarized using "credible" intervals, analogous to traditional confidence intervals. For each posterior distribution in step 4, we determine whether the credible interval for p lies entirely above 0.5.

The probability of excluding the "null" value of 0.5 is then the proportion of simulated data sets for which the credible interval excludes this value. This is an analog to power in a Bayesian context.

Hybrid Prospective Analysis

The previous section described a completely Bayesian prospective (prior predictive) analysis. However, some contexts may require a compromise between the desire to consider prior information and concerns regarding the Bayesian paradigm. In medical research, for example, regulatory authorities do not permit informative priors in post-experimental analysis of clinical data because such priors are considered too subjective (Spiegelhalter, Abrams, & Myles, 2004). Quoting Spiegelhalter and colleagues, DeSantis (2007) wrote that an alternative is "a hybrid frequentist-Bayesian approach 'in which prior information is formally used but final analysis is carried out in a classical framework' " (p. 96). An important advantage of the hybrid approach is that it can make use of valuable prior knowledge for what is essentially a "what if" or sensitivity analysis. This section will discuss two papers illustrating two different approaches to hybrid prospective analysis: Turner, Prevost, and Thompson (2004) and Rotondi and Donner (2009).

Turner et al. (2004), in the medical statistics literature, described power analysis for cluster randomized clinical trials that, like educational data, have a hierarchical structure. They used prior information – published ICC estimates – to inform a classical analysis of power to detect a given effect size difference between treatment and control groups. They compared several different distributional forms for the prior information, which Chapter 5 will discuss in detail. A completely Bayesian analysis would use the prior distribution and a sampling distribution for the data to generate a large number of simulated data sets, and the analysis would

proceed as described in the previous section. However, Turner and colleagues took a "shortcut", using a classical expression for power as a function of between-cluster ICC. They calculated this expression for many random draws from their prior, generating a distribution of power values. In their examples, they found that the mean power calculated from this distribution is lower than the power calculated from a point estimate, because their procedure takes the uncertainty in estimating between-cluster ICC into account.

Rotondi and Donner (2009) presented an empirical Bayes approach to modeling uncertainty in ICC estimation. Rotondi and Donner's priors for ICC, unlike those of Turner et al. (2004), were not fully Bayesian priors using a specific distributional form for the likelihood. Instead, the priors were Gaussian kernel densities fit to the Hedges and Hedberg (2007) summary of ICC estimates for educational data. Kernel density estimation is a well-established method of empirical probability density estimation (e.g., Silverman, 1986; Bowman & Azzalini, 1997), to be discussed in Chapter 5.

Like Turner et al. (2004), Rotondi and Donner performed a hybrid analysis using a prior to directly generate distributions for prospective analysis. Rotondi and Donner considered the number of clusters required for adequate power to detect a given effect size difference between treatment and control groups. They used many random draws from a kernel density fit in a classical formula to generate a distribution for the number of clusters required for a given level of power.

The research methods of this dissertation draw upon and extend the literature reviewed in this chapter. This study innovatively applies the hybrid prospective analysis techniques introduced above to the analysis of power for between-teacher comparisons in VAM. Published between-teacher variance estimates are used to construct both fully Bayesian priors, like those of

Turner et al. (2004), and empirical Bayes priors, like those of Rotondi and Donner (2009). These priors are compared and contrasted, and their implications for power are explored. The next chapter provides a detailed outline of methods.

CHAPTER 3: Method

The ultimate goal of this work is probabilistic description of the power of value-added modeling to detect extreme teacher effects, based on a distribution derived from published between-teacher ICC estimates. This chapter will follow the order of the research questions posed in Chapter 1, describing specific methodologies to address each question.

OpenBUGS (Spiegelhalter, Thomas, Best, Gilks, & Lunn, 2002) provides a language for specifying Bayesian models, and sampling code needed for Markov Chain Monte Carlo (MCMC) simulation. The statistical programming language R (R Foundation for Statistical Computing, 2012) has the power and flexibility needed for all the analyses in the dissertation. The R add-on package R2OpenBUGS **(**Sturtz, Ligges, & Gelman, 2005) facilitates processing of OpenBUGS output.

Review of Published Estimates of Between-Teacher ICC

The first goal of the proposed work is to review recent published between-teacher ICC estimates, since these will be the basis for Bayesian priors on which the rest of the work depends. In particular, it is important to consider the value-added modeling decisions that may have affected the estimates, including:

1. Fixed vs. random (empirical Bayes) estimation of teacher effects
2. Specification of school effects: fixed, random, or covariate adjustment
3. Specification of student effects: fixed or covariate adjustment

Construction and Evaluation of Design Priors

DeSantis (2007) called Bayesian priors incorporating information from previous studies *design priors*. Published ICC estimates will be used to construct the following types of design priors:

1. Explicit distributional assumptions for the likelihood, following Turner et al. (2004). Both of these are large-sample approximations of the standard error of $\hat{\rho}$ assuming normality.
    a. Swiger's distribution
    b. Fisher's distribution
2. Empirical Bayes design priors, following Rotondi and Donner (2009).
    a. Gaussian kernel empirical density
    b. Inverse variance-weighted Gaussian kernel empirical density

The second research question stated in Chapter 1 involves comparing and contrasting these different design priors. One frequently raised objection to Bayesian methods is the subjective judgment required in selecting priors, so it is important to carefully evaluate the implications of prior selection.

Hybrid Power Analysis

Chapter 2 described Schochet and Chiang's (2013) error rate and power calculations for identifying teachers performing significantly above or below average using VAM. Schochet and Chiang used just one estimate of between-teacher ICC and thus failed to account for the uncertainty in estimating this parameter. This study will take a hybrid approach to power calculation for value-added modeling, using a Bayesian design prior to incorporate estimation uncertainty.

Three-Level Hierarchical Model

One consequence of Schochet and Chiang's four-level model was extensive imputation to get point estimates for ICCs (Schochet & Chiang, 2010). This was necessary because in many studies, the data did not permit identification of both a classroom-level and a teacher-level

effect.[4] The proposed work will assume a three-level (students, teachers, and schools) model, requiring only between-teacher and between-school ICC, for which many published estimates are available. The outcome variable g_{ijk} is the gain score for student i of teacher j in school k:

Level 1: Students i = 1, …, n (7)

$$g_{ijk} = \tau_{jk} + \varepsilon_{ijk}, \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$$

Level 2: Teachers j = 1, …, m

$$\tau_{jk} = \eta_k + \theta_{jk},\ \theta_{jk} \sim N\left(0, \sigma_\theta^2\right)$$

Level 3: Schools k = 1, …, s

$$\eta_k = \delta + \psi_k,\ \psi_k \sim N\left(0, \sigma_\psi^2\right)$$

Here δ represents the expected student score gain district-wide. Following Schochet and Chiang and common accountability practice, the HLM will not include student-level or teacher-level covariates.

Design Priors in Power Analysis

Schochet and Chiang (2013) observed that a typical use of VAM "classifies each teacher into a performance category based on the *t* statistic from testing the null hypothesis that the teacher's performance is equal to the average performance in a reference group" (p. 7). That is, the null hypothesis of interest is $H_0: \tau_{jk} - \bar{\bar{\tau}}_{..} = 0$, where $\bar{\bar{\tau}}_{..}$ is the average performance in the district.

From the three-level model of the previous section of equation 7, an expression can be derived for power to detect a teacher whose effect differs from the average by a threshold T. The power β is calculated as follows, where Φ[.] is the normal distribution, T is the threshold value, and model variances $\sigma_\psi^2, \sigma_\theta^2, \sigma_\varepsilon^2$ are described in the previous section:

[4] The studies that did estimate a classroom-level effect did so by comparing a teacher's classrooms over several years, so changes over time in teacher effectiveness would be modeled as classroom-level variance.

$$\beta = \Phi\left[\frac{|T|\sigma}{2\sqrt{V_{est}}}\right] \quad (8)$$

$$\sigma = \sqrt{\sigma_\psi^2 + \sigma_\theta^2 + \sigma_\varepsilon^2} \quad (9)$$

V_{est} is the variance of the teacher effects estimator. For ordinary least squares (OLS) estimation, this variance is:

$$V_{OLS} = \frac{\sigma_\varepsilon^2}{n}\left(\frac{sm-1}{sm}\right) \quad (10)$$

Since variance estimates in the literature are standardized to unit total variance, variances are equal to ICCs, which describe the proportion of total variance from each source (teacher, school, and student). A separate analysis will be conducted for each type of design prior. A large number of values of between-teacher ICC, $\hat{\rho}_T$, will be simulated by drawing values from the design prior distribution. These will be used in equations 8-10 to generate a large number of power values. The result will be a distribution for power that is derived from the design prior distribution of $\hat{\rho}_T$ – in other words, probabilities of achieving various levels of power, based on the assumed distribution of $\hat{\rho}_T$.

Other Simulation Parameters

The proposed work will adopt other simulation parameter values from Schochet and Chiang (2013), including sample sizes n = 21 students per class, m = 10 teachers per school, and s = 5 schools per district. Schochet and Chiang used a between-school ICC of 0.005, which, like their between-teacher ICC, is the mean of values from recent studies.

Schochet and Chiang set the threshold T to represent an educationally meaningful difference of a teacher from the average, as described in Chapter 2. Their between-teacher ICC estimate was an average of estimates from both math and reading test scores, and their analysis

was intended to apply to VAM from either type of score. The subject-specific analysis described in this dissertation will set different thresholds for math and reading, calculated from the different average annual growth estimates for the subjects (0.94 standard deviations for math and 0.65 standard deviations for reading). Schochet and Chiang's medium threshold value was 0.2 standard deviations, representing 21% of an average annual gain for math and 31% for reading. Here, the medium threshold will represent 25% of an average annual gain for each subject. The low threshold will be half the medium, and the high threshold 1.5 times the medium, following Schochet and Chiang. For example, the medium math threshold will be 0.25*0.94 = 0.24 standard deviations, the low threshold 0.12 standard deviations, and the high threshold 0.36 standard deviations. The values for reading will be 0.08, 0.16, and 0.24 standard deviations.

The next chapter begins the technical description of the research with the review of published between-teacher ICC estimates.

Chapter 4: Between-Teacher ICC Estimates from the VAM Literature

This study's key innovation is the use of Bayesian probability distributions to represent empirical evidence for various values of between-teacher ICC. This chapter will survey recent between-teacher ICC estimates from the VAM literature. These will be used in Chapter 5 to generate Bayesian design priors for power analysis of VAM. The review will explore central issues for structuring the design prior analysis:

1. Are there systematic differences between estimates calculated using different model specifications?
2. Are there systematic differences between estimates from math and reading test scores?
3. As a preliminary approach to summarizing estimates, what does classical meta-analysis indicate about their distribution?

Table 1 summarizes between-teacher variance estimates for U. S. elementary teachers from VAM studies of math and reading test scores published in peer-reviewed journals in the last ten years. Since the score data in these papers are standardized, the estimates are also intra-class correlations (ICCs).[5] The fourth column indicates the statistical method used to estimate between-teacher variance. The next section will describe and compare the different methods used by these authors, and also discuss other model specifications affecting the estimates.

Two recent papers provide estimates involving substantial methodology differences from those in Table 1. The method of Rivkin, Hanushek, and Kain (2005) "makes use of information on teacher turnover and grade average achievement gains to generate a lower bound estimate of the within-school variance in teacher quality" (p. 425). They reported estimates of 0.012 for

[5] Rockoff (2004) used a nationally standardized scale, unlike the other authors in Table 1, who scaled at the district level. Jacob and Lefgren (2005) commented on Rockoff's relatively small estimates: "0.1 standard deviations on a national scale (which Rockoff uses) may be much more than 0.1 standard deviations within the distribution of achievement within Rockoff's districts. This would be true if the students in his district were substantially more homogenous than is true nationally" (p. 15).

math and 0.010 for reading. As "lower bounds," these are small compared to those in Table 1, especially the math estimate. Plecki, Elfers, and Nakamura (2012) estimated VAMs without a school-level component, so their between-teacher variance includes between-school variation, and their estimates are large compared to those in Table 1 (0.11 for math and 0.07 for reading). In the interest of methodological consistency, Rivkin et al. (2005) and Plecki et al. (2012) are excluded from this study.

Table 1

Teacher Effects Variance Estimates

Study	Location	Grades	Variance Estimation Method	Subject	Teacher Effects Variance/ICC
Jacob & Lefgren (2008)	Western US school district	2-6	Variance of fixed effects (sampling error adjusted)	Math	0.068
				Reading	0.014
Kane, Rockoff, & Staiger (2008)	New York City	4-8	Covariance of fixed effects from different years	Math	0.017
				Reading	0.010
Kane & Staiger (2008)	Los Angeles Unified School District	2-5	Covariance of fixed effects from different years	Math	multiple
				Reading	multiple
Koedel & Betts (2011)	San Diego Unified School District	4	Variance of fixed effects (sampling error adjusted)	Math	0.048
Nye et al. (2004)	TN	K-3	Variance of random effects	Math	multiple
				Reading	multiple
Rockoff (2004)^	NJ (two districts)	K-6	Variance of fixed effects (sampling error adjusted)	Math	0.011
				Reading	0.009
Rothstein (2010)*	NC	3-5	Variance of fixed effects (sampling error adjusted)	Math	0.023
				Reading	0.012
Sass, Hannaway, Xu, Figlio, & Feng (2012)	FL, NC	3-5	Variance of random effects	Math	multiple
				Reading	multiple

^ Estimates were averaged from Table 2 in Rockoff (2004), weighted by number of teachers.
***** Two estimates from Rothstein (2010) are excluded because the fixed effects variance was not adjusted for sampling error, so these estimates are not comparable with the others in the table.

Model Specifications Affecting Between-Teacher ICC Estimates

Two studies in Table 1 (Nye et al., 2004 and Sass et al., 2012) estimate teacher effects as normally distributed random effects, $\theta \sim N\left(0, \sigma_\theta^2\right)$, where σ_θ^2 is the between-teacher variance. Kane and Staiger (2008) wrote that the idea of this approach "is to multiply a noisy estimate of teacher value added… by an estimate of its reliability, where the reliability of a noisy estimate is the ratio of signal variance to signal and noise variance" (p. 14). This is sometimes called *shrinkage* because individual teacher estimates are adjusted towards the overall mean estimate. Shrinkage corrects for the greater sampling error in small samples (Gelman & Hill, 2007).

Lockwood and McCaffrey (2007) described a potential problem with random effect estimates: "Correlation between the unobserved individual effects and other substantive variables in the model can lead to biased and inconsistent estimates of the effects of those variables" (p. 225). Such correlation could arise if students are not randomly assigned to teachers – a topic of intense debate in the VAM literature. In a widely discussed study, Rothstein (2010) found that fifth grade teacher assignments predicted third and fourth grade score gains. "Since teachers cannot rewrite the past," as the National Research Council and National Academy of Education (2010) explained (p. 49), this result suggests that fifth grade teaching assignments are correlated with earlier achievement gains – in other words, that students are not randomly assigned to fifth grade teachers.

Researchers concerned about nonrandom assignment often prefer to treat teacher effects as fixed effects – that is, indicator or "dummy" variables for individual teachers. This is the most common variance estimation method in Table 1. When estimating between-teacher variance, researchers using fixed teacher effects adjust the variance of the effects *post hoc* for sampling error. The closely related method of Kane and colleagues (Kane et al., 2008; Kane & Staiger,

2008) estimates between-teacher variance as the covariance of fixed teacher effects from different school years.

Teacher effect specification is not the only modeling decision affecting the between-teacher variance estimates. Specification of student and school effects impacts estimation of between-school and between-student (residual) variances, which in turn affect the relative contribution of between-teacher variance to total variance (ICC). Like teacher effects, school effects may be specified as random; however, this raises concern regarding nonrandom assignment of students and teachers to schools. Nye et al. (2004) is the only paper in Table 1 to estimate school effects as random effects. As with teacher effects, an alternative is to treat school effects as fixed, like Jacob and Lefgren (2008), Kane and Staiger (2008), and Rothstein (2010).

School fixed effects have an important disadvantage from the accountability perspective: teachers can be compared to other teachers in the same school, but teachers cannot be compared across schools (Baker et al., 2010). An alternative model specification instead controls for school-level covariates believed to affect test score outcomes. Kane et al. (2008) controlled for school average class size and school means of student-level covariates, while Sass et al. (2012) controlled for principal experience and other unspecified covariates. Koedel and Betts (2011) used both school fixed effects (representing characteristics stable over time) and time-varying school-level covariates, including class size and percentage of students who were English language learners and who qualified for free/reduced price lunch.

Besides specifying teacher and school effects, researchers must decide how to control for differences in students' preparation for learning. Models used in accountability practice generally assume that past test scores alone adequately represent student background (Ballou, Sanders, & Wright, 2004). All the studies in Table 1 controlled for previous test scores, and Nye et al.

(2004) and Rothstein (2010) used no other student background controls. Most researchers, however, used either additional student-level covariates, or student fixed effects. The choice is analogous to that between school fixed effects or school-level covariate adjustment, described above.

Koedel and Betts (2011) described their rationale for using student fixed effects: "Econometric theory suggests that the inclusion of student fixed effects will be an effective way to remove within-school sorting bias in teacher effects as long as students and teachers are sorted based on time-invariant characteristics" (p. 29). Rockoff (2004) also used student fixed effects. On the other hand, Jacob and Lefgren (2008) and Kane et al. (2008) chose to control for student-level covariates, such as race, gender, and socioeconomic status. Kane and Staiger (2008) and Sass et al. (2012) reported estimates from models incorporating student fixed effects, and also from other models incorporating student covariates. Sass and colleagues noted that their covariate models possibly omitted variables that influenced between-student variation, while student fixed effects captured variation from both observed and unobserved variables.

Comparisons of Estimates from Recent Literature

Figure 1 shows the estimates from Table 1, using symbols to denote different teacher effect estimation methods. There are obvious differences between estimates from math test scores (top panel) and reading test scores (bottom panel): math estimates tend to be larger than reading estimates, and their spread is also larger. This intriguing finding has been little discussed in the literature. Nye et al. (2004) speculated, "This may be because mathematics is mostly learned in school and thus may be more directly influenced by teachers, or that there is more variation in how (or how well or how much) teachers teach mathematics. Reading, on the other hand, is more likely to be learned (in part) outside of school and thus the influence of school and

teacher on reading is smaller, or there is less variation in how (or how well or how much) reading is taught in school" (p. 247). However, it is also possible that measurement issues, arising from differences in the instruments used to measure math and reading achievement, contribute to the differences in estimates. This idea will be further discussed in the final chapter.

Figures 2 and 3 illustrate that there are also some systematic differences between estimates calculated using different model specifications. Figure 2 shows the school effect specifications used for each study. A striking pattern in Figure 2 is the much larger between-teacher variance estimates of Nye et al. (2004), the only estimates from models using school random effects. Random-effects estimates are "shrunk" to the overall mean as described above, so the between-school variance estimates are small relative to fixed effects. This is especially true for these authors' school random effects estimates, because they are based on small samples of classrooms within schools, sometimes as few as four classrooms. A smaller between-school variance implies that between-teacher variation becomes a larger proportion of total variance – in other words, between-teacher ICC is larger, relative to fixed effects.

Figure 3 shows the student effect specifications used for each study. Estimates tend to be larger in models with student covariate adjustment (circles) than in models with student fixed effects (triangles). As Sass et al. (2012) pointed out, covariate adjustment takes into account only the variables specified, so it will fail to account for variation between students due to variables not included in the model. Fixed effects will model more variation as between-student variation, which may reduce the amount of variation identified as between-teacher variation. In Kane and Staiger's (2008) interpretation, "differencing out student fixed effects in test score levels understates teacher differences" (p. 3).

Meta-Analysis of ICC Estimates

Between-teacher ICC can be interpreted as the correlation of outcomes for students of the same teacher, so classical methods of meta-analysis for correlations offer a preliminary approach to describing the distribution of the estimates. Classical meta-analysis will serve as a reference point for interpreting and "sanity checking" the complex Bayesian distributions in the next two chapters. As discussed in the last section, the estimates of Nye et al. (2004) were the only recent estimates involving random school effects, and these differed substantially from the other studies that used fixed school effects or school-level covariates. The Nye et al. (2004) estimates have been excluded from the meta-analysis.

As described in Hartung, Knapp, and Sinha (2008), a combined estimate of the correlation is a combination of the study estimates weighted by their asymptotic variance estimates $\widehat{Var}(r_i) = \frac{\left(1-r_i^2\right)^2}{n_i-1}$:

$$\tilde{\theta} = \frac{\sum r_i/\widehat{Var}(r_i)}{\sum 1/\widehat{Var}(r_i)}, \widehat{Var}\left(\tilde{\theta}\right) = \frac{1}{\sum 1/\widehat{Var}(r_i)} \tag{11}$$

A chi-squared statistic for testing homogeneity of estimates is:

$$\chi^2 = \sum \frac{\left(r_i-\tilde{\theta}\right)^2}{\widehat{Var}(r_i)} \tag{12}$$

The homogeneity hypothesis H_0: $\theta_1 = \ldots = \theta_k$ is rejected if $\chi^2 > \chi^2_{k-1,\alpha}$.

The results are shown in Table 2. Note that the 95% CI for reading (just) includes 0, supporting the possibility of no correlation of reading scores between students of the same teacher. The test for homogeneity of correlations retains the null hypothesis that all the observations are estimating a common underlying population correlation, despite the diversity of geographical areas from which the estimates come. It is important to note that the available estimates are not geographically representative of the U. S. With the possible exception of Jacob

and Lefgren (2008), the states and districts represented are all on the east and west coasts. States and districts with data collection and management processes adequate to support VAM studies might plausibly differ in important ways from areas not yet studied.

In next chapter, the published estimates of between-teacher ICC reported in Table 1 will be used to construct fully Bayesian and empirical Bayes distributions representing uncertainty in estimating this parameter. The review in this chapter informs two initial decisions about these design priors. First, the estimates of Nye et al. (2004) will be excluded for the reasons discussed above. Second, separate priors will be constructed for estimates from math and reading test scores. Although the classical meta-analysis suggests that all the observations are estimating a common underlying parameter, the differences between subjects appear to be of sufficient statistical and practical interest to treat separately.

Table 2

Results of Classical Meta-Analysis for Correlations

	All	**Math**	**Reading**
Estimate of common population correlation	0.017	0.024	0.010
95% CI for common population correlation	(0.009, 0.024)	(0.013, 0.034)	(-0.001, 0.021)
Test for homogeneity of correlations	$\chi^2 = 5.439$	$\chi^2 = 1.775$	$\chi^2 = 0.467$
H_0: $\rho_1 = \rho_2 = \cdots = \rho_k$	$\chi^2_{0.05,14} = 23.685$	$\chi^2_{0.05,7} = 14.067$	$\chi^2_{0.05,6} = 12.592$
	Retain H_0	Retain H_0	Retain H_0

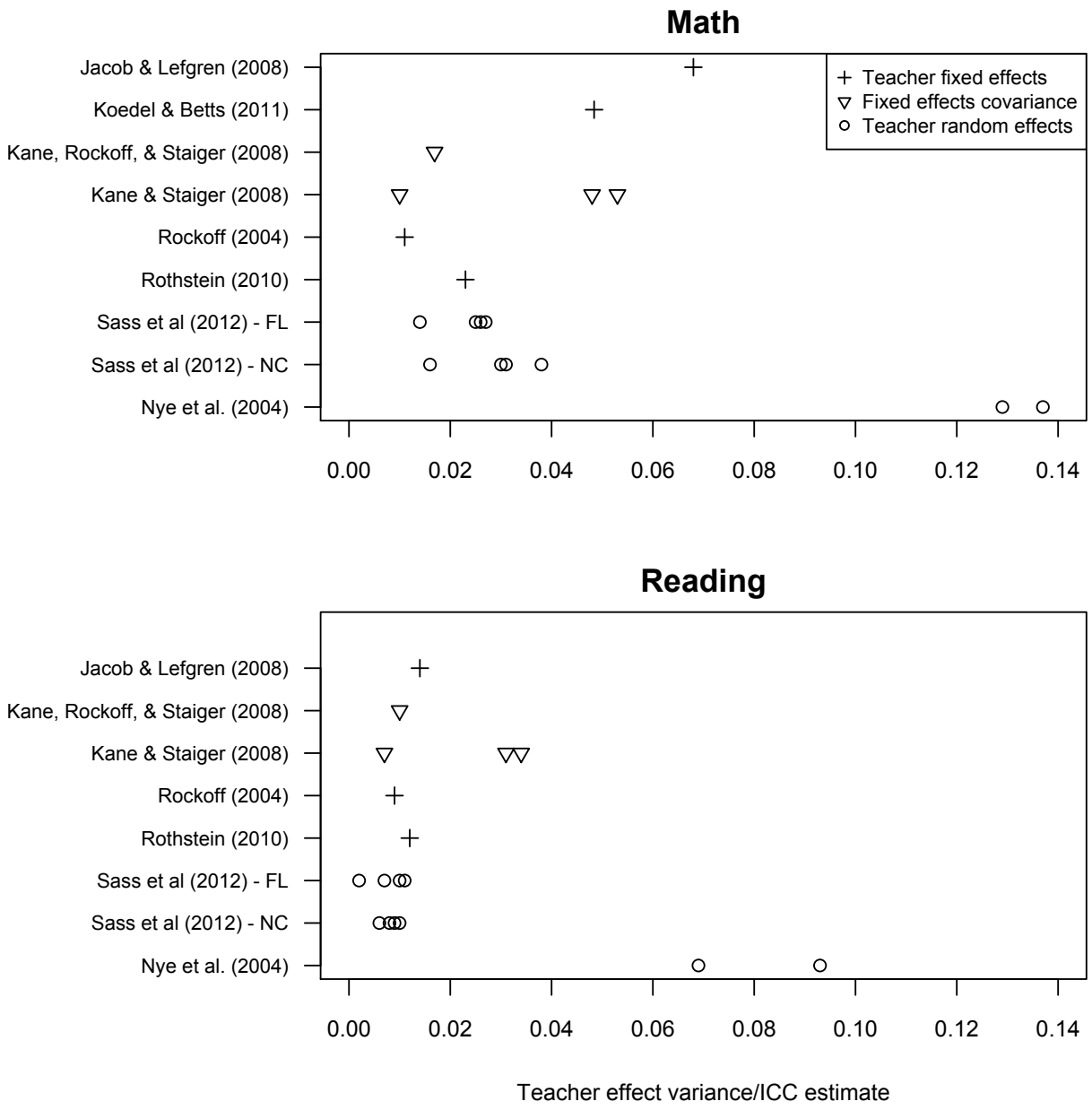

Figure 1. Teacher effect estimates and methods of estimation from recent studies.

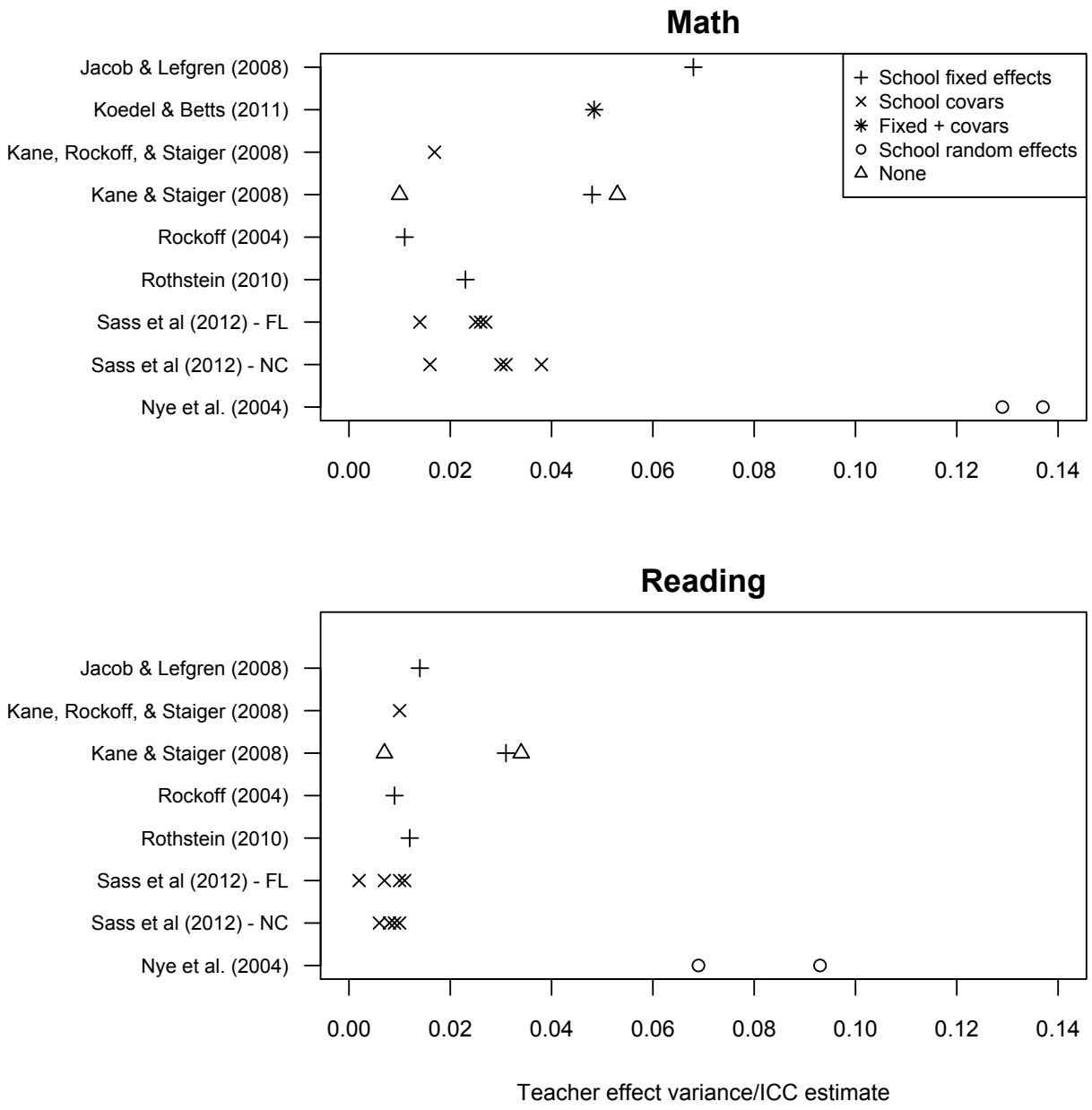

Figure 2. Teacher effect estimates from recent studies, with specifications of school effects.

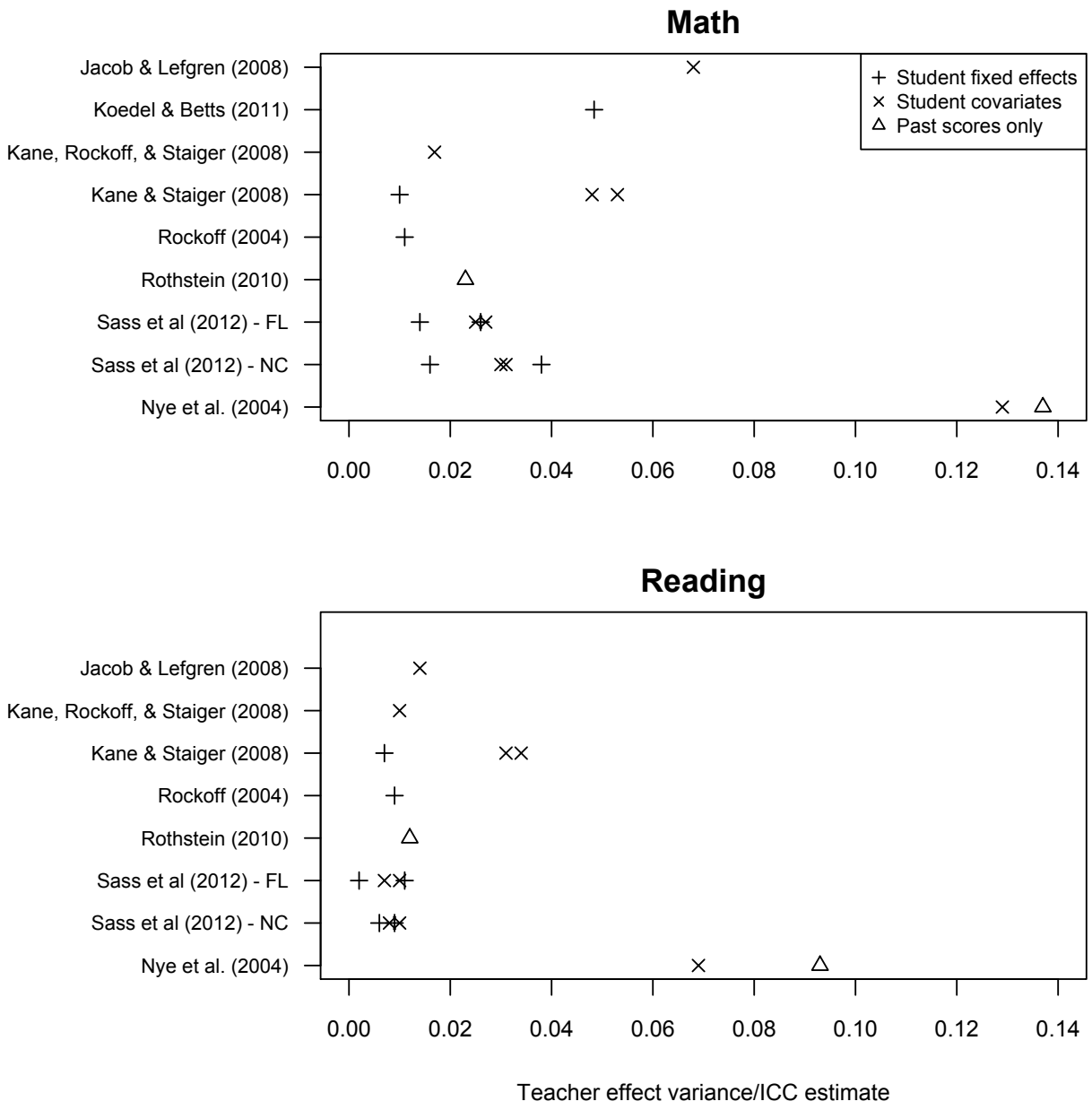

Figure 3. Teacher effect estimates from recent studies, with specifications of student effects.

CHAPTER 5: Describing Uncertainty in Estimating Between-Teacher ICC

As VAM has gained attention from researchers and policy makers, US states and districts have begun using VAM for school and teacher accountability purposes (National Research Council and National Academy of Education, 2010). These uses of VAM imply an assumption that VAM estimates of teacher effects can reliably differentiate higher- from lower-performing teachers. In fact, this differentiation depends critically on the proportion of the overall variation in test scores that can be attributed to differences between teachers – that is, the intra-class correlation (ICC). Schochet and Chiang (2013) observed that a typical use of VAM "classifies each teacher into a performance category based on the *t* statistic from testing the null hypothesis that the teacher's performance is equal to the average performance in a reference group" (p. 7). The ability to correctly classify teachers can be thought of in terms of statistical power (probability of correctly rejecting the null, or 1 - Type II error rate). The power the test can achieve depends on the intra-class correlation: the larger the between-teacher variation relative to the other sources of variation, the greater the power to detect differences between teachers.

Researchers frequently conduct prospective analysis prior to collecting data to evaluate whether their design will achieve study goals (such as reasonable power). Turner et al. (2004) noted that in the medical statistics literature, prospective power calculations had typically used ICC estimates without taking into account the uncertainty in their estimation. The same has been true of prospective analysis for VAM, as in the case of a recent example, Schochet and Chiang (2013). As discussed in Chapter 2, these authors performed prospective analysis of error rates for an HLM of test score gains with four levels (students, classrooms, teachers, and schools). From the normality assumptions of classical hypothesis testing, they derived expressions for Type I and II error rates for identifying teachers and schools performing significantly above or below

average, as well as expressions for overall misclassification rates for the entire system. They then evaluated these expressions for plausible values of variances and sample sizes.

Schochet and Chiang derived their variances from point estimates of ICC, calculated as means of ICCs from 15 recent studies. They performed minimal analysis of the sensitivity of their results to alternative ICC values. As demonstrated in Chapter 4, published estimates of between-teacher ICC show considerable variation. Thus the uncertainty in estimating between-teacher ICC – that is, the range of possible values supported by published estimates – demands attention in prospective analysis. This chapter will describe Bayesian approaches to quantifying uncertainty about between-teacher ICC. Bayesian methods have become well-established strategies for describing parameter uncertainty, because they treat information about model parameters as probability distributions (Gelman & Hill, 2007). These distributions provide a much fuller description of the plausibility of parameter values than point estimates or confidence intervals.

ICC Uncertainty for Prospective Analysis: A Bayesian Approach

Chapter 2 introduced the concept of a Bayesian posterior distribution, derived from a prior distribution and the data likelihood:

$$p(\theta|y) \propto p(y|\theta)\pi(\theta) \tag{13}$$

where $p(y \mid \theta)$ is the data likelihood. A prior distribution $\pi(\theta)$ is specified for the parameters. In empirical Bayes estimation, the observed data distribution serves as the prior. Fully Bayesian approaches, on the other hand, allow the analyst to incorporate truly prior information – knowledge or beliefs about parameter distributions, held before data is observed.

In cases where little is known about the parameters, the prior $\pi(\theta)$ may have a "non-informative" distribution in which all possible values have the same probability of occurrence.

However, for prospective analysis, the prior typically incorporates information from previous studies; DeSantis (2007) called such priors *design priors*. The design prior "updates" a noninformative initial prior $\pi_0(\theta)$ with information from previous studies, **z**, expressed as a likelihood $L(\theta; \boldsymbol{z})$:

$$\pi^D(\theta|\boldsymbol{z}) \propto \pi_0(\theta)L(\theta;\boldsymbol{z}) \quad (14)$$

Specific likelihood forms used here will appear in the following sections, which will describe design prior generation strategies from two recent papers introduced in Chapter 2. Turner et al. (2004) took a fully Bayesian approach, specifying distributional forms for $\pi_0(\theta)$ and $L(\theta; \boldsymbol{z})$. Rotondi and Donner (2009), on the other hand, used an empirical Bayes strategy, generating design priors directly from the observed data distribution.

Fully Bayesian Design Priors

Turner et al. (2004), in the medical statistics literature, described power analysis for cluster randomized clinical trials that, like educational data, have a hierarchical structure. Their hybrid approach was developed in the context of medical research, where regulatory authorities do not permit informative Bayesian priors in final analysis because such priors are considered too subjective (Spiegelhalter et al., 2004). Turner and colleagues therefore used a design prior for the within-cluster ICC (ρ) to inform a classical power analysis.

Their initial prior $\pi_0(\rho)$ for ρ was Uniform[0,1]. They compared the use of several different distributional forms for the likelihood $L(\rho; \boldsymbol{z})$, resulting in several versions of the design prior $\pi^D(\rho|\boldsymbol{z})$. The discussion of likelihoods in this section will assume that a single underlying ρ is common to all studies. An alternative framework would be a hierarchical structure in which different underlying values are assumed for each study (ρ_m, m = 1 … r, where r is the number of studies), and the distribution of the ρ_m is modeled. In the language of meta-analysis, this

distinction corresponds to "fixed-" vs. "random-effects" meta-analysis of the available estimates (Hartung et al., 2008). As noted in Chapter 3, for both math and reading, the test for homogeneity of correlations traditionally used in meta-analysis retains the null hypothesis of a common underlying population correlation. Also, while the fully Bayesian approach to be described is sufficiently flexible to implement either a fixed- or random-effects meta-analysis, the empirical Bayes design priors to be described below implicitly assume that a common ρ is being estimated. For comparability with the empirical Bayes design priors, the following discussion of Turner and colleagues' likelihood forms will be in the context of estimating a single common ρ.

From the work of Ukoumunne (2002), Turner and colleagues selected likelihood forms for *L(ρ; z)* that can be computed from minimal information about the studies providing the estimates: total sample size and number of clusters. Two of the likelihood forms (Swiger's distribution and Fisher's distribution) are large sample approximations assuming normality, while the third (Searle's distribution) is based on the F distribution of the variance ratio in one-way ANOVA.

1. *Swiger's distribution.* Swiger assumed normality of the estimates $\hat{\rho}_m$:

$$\hat{\rho}_m \sim N\left(\rho, Var(\hat{\rho}_m)\right), Var(\hat{\rho}_m) = \frac{2(N_m - 1)(1-\rho)^2\left\{1 + \left(\frac{N_m}{k_m} - 1\right)\rho\right\}^2}{\left(\frac{N_m}{k_m}\right)^2 (N_m - k_m)(k_m - 1)} \quad (15)$$

where m = 1 … r studies, N_m is the total number of observations in study m and k_m is the number of clusters in study m, so that N_m / k_m represents average cluster size for study m.

2. *Fisher's distribution.* Fisher assumed normality for a transformation of $\hat{\rho}_m$:

$$g(x) = \frac{1}{2}\log\left[\frac{1 + (N_m/k_m - 1)x}{1 - x}\right] \quad (16)$$

$$g(\hat{\rho}_m) \sim N\left(g(\rho), \frac{1}{2}\{(k_m - 1)^{-1} + (N_m - k_m)^{-1}\}\right)$$

3. *Searle's method.* This distribution is based on the variance ratio statistic in one-way ANOVA:

$$\left\{\frac{1+(N/k-1)\hat{\rho}_m}{1-\hat{\rho}_m}\right\}\left\{\frac{1-\rho}{1+(N/k-1)\rho}\right\} \sim F_{k-1,N-k} \quad (17)$$

Empirical Bayes Design Priors: Gaussian Kernel Densities

Rotondi and Donner (2009) presented an empirical Bayes approach to modeling uncertainty in ICC estimation. Rotondi and Donner's design priors, unlike Turner, Prevost, and Thompson's, were not combinations of a prior and a distributional form for the likelihood. Instead, $\pi^D(\rho|z)$ was derived in one step as a Gaussian kernel density fit to Hedges and Hedberg's 2007 summary of ICC estimates for educational data. Rotondi and Donner generated both an unweighted distribution and a distribution from the estimates weighted by the inverses of their estimated variances, to account for the increased precision of estimates from larger samples.

Kernel density estimation is a well-established method of empirical probability density estimation (e.g., Bowman & Azzalini, 1997; Silverman, 1986). For data $X_1, X_2, \ldots, X_n$ treated as a sample from a random variable with density *f,* the kernel density estimate of *f* is

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \quad (18)$$

The kernel function K satisfies $\int K(x)dx = 1$ and is often chosen to be Gaussian (normal) in shape. The method can be thought of as constructing a Gaussian curve around each estimate, then summing the densities at each value in the range to obtain an overall distribution. The parameter *h,* known as the bandwidth, controls the smoothness of the estimated density function. As Sheather (2004) noted, "It is well known that the value of the bandwidth is of critical importance, while the shape of the kernel function has little practical impact" (p. 589).

Kernel density estimation is implemented in the base distribution of R (R Foundation for Statistical Computing, 2012).

Interpreting Fully Bayesian and Empirical Bayes Design Priors

Two different sources of variability are involved in generating design priors from the published estimates: first, uncertainty about the underlying parameter value, and second, sampling variability in the estimates. Although the test of homogeneity in Chapter 4 does not reject the null, it would be difficult to make a reasonable *a priori* argument that the between-teacher ICC is precisely the same everywhere at all times. Thus there is uncertainty about the underlying parameter value. On top of this, the data sets used in the studies are samples from the population of US elementary schools, teachers, and students, and are thus subject to sampling variability. In other words, even if the underlying parameter value were exactly the same across time and geography, the estimates would still vary with the schools, teachers, and students selected.

A fully Bayesian design prior with a specific distributional form for the likelihood explicitly separates the two types of variability. Parameter uncertainty is represented as a distribution expressing a vague prior belief about ρ, such as Uniform[0,1]. This is combined with the data using a likelihood function representing the sampling distribution. For example, in the Swiger and Fisher likelihoods, sampling variability is represented as $Var(\hat{\rho}_m)$, calculated as shown in equations 15 and 16. By contrast, the empirical Bayes estimates do not separate parameter uncertainty and sampling variability; the spread of the distribution includes both.

This distinction has implications for using and interpreting fully Bayesian and empirical Bayes design priors. For example, a fully Bayesian approach makes it straightforward to incorporate new information (in this case, newly published estimates) using Bayesian updating.

The design prior simply becomes the initial prior $\pi_0(\theta)$, and the same distributional form for the likelihood is used to generate a new design prior that incorporates the new estimates. Empirical densities, on the other hand, cannot be updated in the same way; incorporating new estimates would require a new fit. Also, the sampling variability included in empirical densities raises concerns about overfitting to estimates from small studies in the tails of the distribution (Silverman, 1986).

Although a fully Bayesian approach has important advantages, Rotondi and Donner (2009) raised the issue that the distribution of published estimates may not be well described by commonly used probability distributions. They favored an empirical Bayes approach because it "allows the prior distribution to be obtained using all the data at hand, without any particular distributional assumptions" (p. 235). One specific concern for this work is whether the Swiger and Fisher likelihoods, developed for randomized studies, make sense for VAM observational studies.

The rest of the chapter describes construction of design priors for between-teacher ICC from published estimates. Properties of the resulting design priors will be described, and priors generated by different methods will be compared.

Design Priors from Published Estimates of Between-Teacher ICC

Published Estimates Included in Analysis

The estimates used for constructing design priors are presented in Table 3. Sass et al. (2012) estimated between-teacher ICC (ρ) separately for Florida and North Carolina, and the two states are treated as two separate studies here. Two studies provided multiple estimates of between-teacher ICC for both math and reading: Kane & Staiger (2008) and Sass et al. (2012).

The multiple estimates were averaged to provide a single estimate from each study (shown in Table 3)[6].

Fully Bayesian Design Priors

The published estimates of between-teacher ICC discussed in Chapter 4 were treated as estimates of a single underlying ("fixed-effects") between-teacher ICC. Estimation using the Swiger and Fisher likelihoods was performed using OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) and R (R Foundation for Statistical Computing, 2012). Code appears in the Appendix. Convergence was checked using the Gelman-Rubin statistics (Lunn et al., 2013). Estimation was performed separately for math (eight estimates) and reading (seven estimates).

Table 3

Estimates Used for Constructing Design Priors

Study	Location	k (number of teachers)	N (number of students)	Subject	Teacher Effects Variance/ICC
Jacob & Lefgren (2008)	Western US school district	201	20,100*	Math	0.068
				Reading	0.014
Kane et al. (2008)	New York City	11,300^	374,000^	Math	0.017
		11,400^	364,500^	Reading	0.010
Kane & Staiger (2008)	Los Angeles	1,950	43,766	Math	0.037
				Reading	0.024
Koedel & Betts (2011)	San Diego	389	15592	Math	0.048
Rockoff (2004)	NJ	263	24,705	Math	0.011
		224	23,921	Reading	0.009
Rothstein (2010)	NC	3,040	60,740	Math	0.023
				Reading	0.012
Sass et al. (2012)	FL	9,170	733,600+	Math	0.023
		9,396	751,680+	Reading	0.008
Sass et al. (2012)	NC	7,965	637,200+	Math	0.029
		7,957	636,560+	Reading	0.008

*Not specified in paper, estimated as 201 teachers x 20 students/teacher x 5 years.
^ N and k for elementary estimates not specified in paper. Estimated as 0.6 x total N and total k (the study included grades 4-8, of which grades 4-6 were elementary school).
+ Not specified in paper, estimated as k x 20 students/teacher x 4 years.

[6] Estimates from Sass et al. (2012) were also averaged across poverty condition (high and low).

It became apparent that the Searle likelihood was infeasible for studies of the sizes typical in value-added modeling. The reason is evident from the F distribution likelihood:

$$p(x|n,m) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} x^{\frac{n}{2}-1} \left(\frac{n}{m}\right)^{\frac{n}{2}} \left\{1 + \frac{nx}{m}\right\}^{\frac{-(n+m)}{2}} \quad (19)$$

where n = k – 1, m = N – k. Searle derived this result with an eye to the analysis of moderately-sized randomized controlled trials (Ukomunne, 2002). For the large N and k in Table 3, the values of the gamma function (continuous extension of factorial) and the powers in the second half of the equation attain very large or small values that software treats as infinity or zero.

Table 4 summarizes distribution of between-teacher ICC ρ generated from the Swiger and Fisher likelihoods. The 2.5 and 97.5 percentiles correspond to a 95% credible interval, a Bayesian analog to a 95% confidence interval. Figure 4 shows the distributions as histograms of simulated ρ values observed across 2700 simulations.

Table 4

Summary of Design Priors from Swiger and Fisher Likelihoods

	percentiles of ρ		
	2.5%	50%	97.5%
Math estimates			
Swiger	0.025	0.026	0.026
Fisher	0.023	0.024	0.024
Reading estimates			
Swiger	0.008	0.009	0.009
Fisher	0.008	0.008	0.009

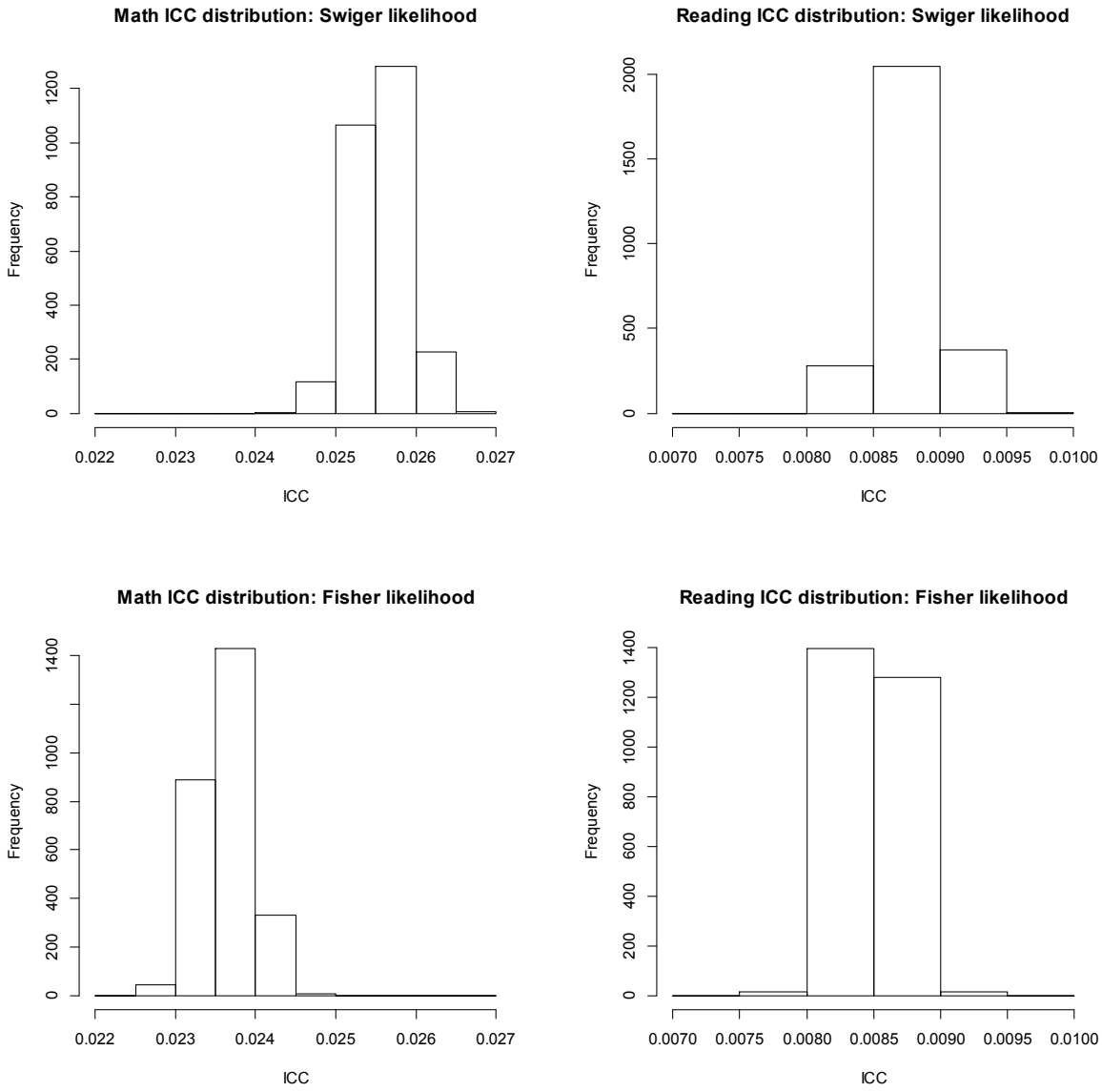

Figure 4. Histograms of simulated ρ values from 2700 simulations.

The deviance information criterion (DIC) is often used to compare the fit of different models estimated using MCMC (Lunn et al., 2013). Unfortunately, the DICs for the Swiger and Fisher likelihoods are not comparable, because the Fisher likelihood involves a data transformation. An approach to assessing model fit often used in Bayesian analysis is "to simulate replicated datasets from the fitted model and compare these to the observed data"

(Gelman & Hill, 2007, p. 513). The approach is described below, using the specific example of math estimates with the Swiger method.

1. A random value of ρ was drawn from the set of simulated ρ values shown in the upper left histogram in Figure 3.1. This represented the mean of the distribution of estimates, $\hat{\rho}_m \sim N\left(\rho, Var(\hat{\rho}_m)\right)$.
2. To represent the variances of the observed estimates, $Var(\hat{\rho}_m)$ (m = 1,…, 8) were calculated as $Var(\hat{\rho}_m) = \frac{2(N_m-1)(1-\rho)^2\left\{1+\left(\frac{N_m}{k_m}-1\right)\rho\right\}^2}{\left(\frac{N_m}{k_m}\right)^2(N_m-k_m)(k_m-1)}$, where ($N_m$, k_m) were the values from the eight published studies used.
3. Given the values from steps 1 and 2, $\hat{\rho}_m$ (m = 1,…, 8) were drawn from $\hat{\rho}_m \sim N\left(\rho, Var(\hat{\rho}_m)\right)$.
4. Steps 1-3 were repeated to generate 10,000 data sets, each representing a simulated set of eight published math estimates. Each of the 10,000 sets was summarized as a five-number summary (minimum, 1st quartile, median, 3rd quartile, maximum).
5. The 10,000 minima were summarized as a five-number summary, and the same was done for the quartiles and the maxima.

Figure 5 shows minima, 1st quartiles, medians, 3rd quartiles, and maxima of the observed data (sets of published estimates) as open circles. The distributions of the corresponding values from 10,000 simulated data sets are shown as boxplots.

The maximum published ICC estimate for reading, from Kane et al. (2008), is much larger than the largest simulated maximum from the Swiger method. However, the quantiles of published reading estimates fall within the ranges of the simulated values from the Fisher method (with the exception of the first quartile). This suggests that the Fisher likelihood is a better fit to

the reading estimates than the Swiger likelihood. Figure 4 shows that the results from the two different likelihoods are so similar that the choice may have little practical impact for this study.

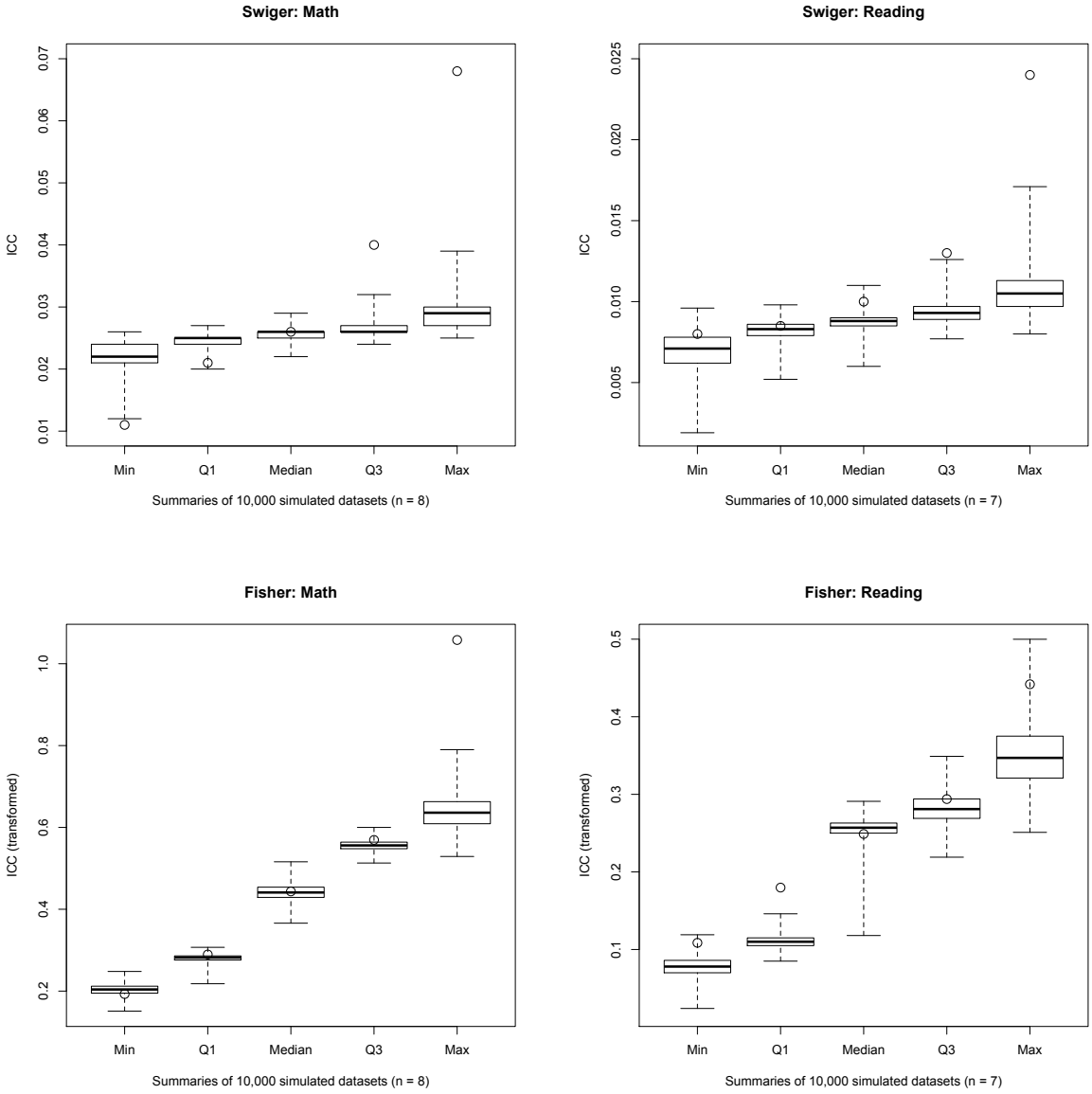

Figure 5. Five-number summaries of published estimates (open circles), compared with 10,000 simulated data sets (boxplots). (Note that the Fisher ICCs are transformed as $g(x) = \frac{1}{2}\log\left[\frac{1+(N_m/k_m-1)x}{1-x}\right]$.)

The maximum published ICC estimate for math, Jacob and Lefgren (2008), is much larger than the largest simulated maxima from both the Swiger and Fisher likelihoods. T he

minimum estimate (Rockoff, 2004) falls below the Swiger minima, suggesting that the Fisher likelihood is a better fit to the math estimates than the Swiger likelihood. The histograms in Figure 4 show that the Fisher estimates for math are somewhat smaller than the Swiger estimates.

Empirical Bayes Design Priors

Gaussian kernel density estimates were generated using the density() function in R (R Foundation for Statistical Computing, 2012). Estimation was performed separately for math (eight estimates) and reading (seven estimates). Bandwidth was initially determined using the widely recommended Sheather-Jones method (e.g., Bowman & Azzalini, 1997).

Figure 6 shows the density estimates. For the math estimates, the Sheather-Jones (SJ) bandwidth results in a density with a peak in the right tail from the Jacob and Lefgren (2008) estimate. The reading density estimate also has a peak in the right tail from Kane et al. (2008). Figure 6 compares the cumulative density estimates with the empirical cumulative densities, a method of checking fit (Loader, 1999). For the reading estimates, the cumulative density estimate corresponds well with the empirical cumulative density; for the math estimates, on the other hand, it falls below the empirical cumulative density, indicating that the peak heights are being underestimated. Inspection of several bandwidths revealed that a kernel density estimate with bandwidth of 0.75 times the SJ bandwidth best tracks the empirical cumulative density.

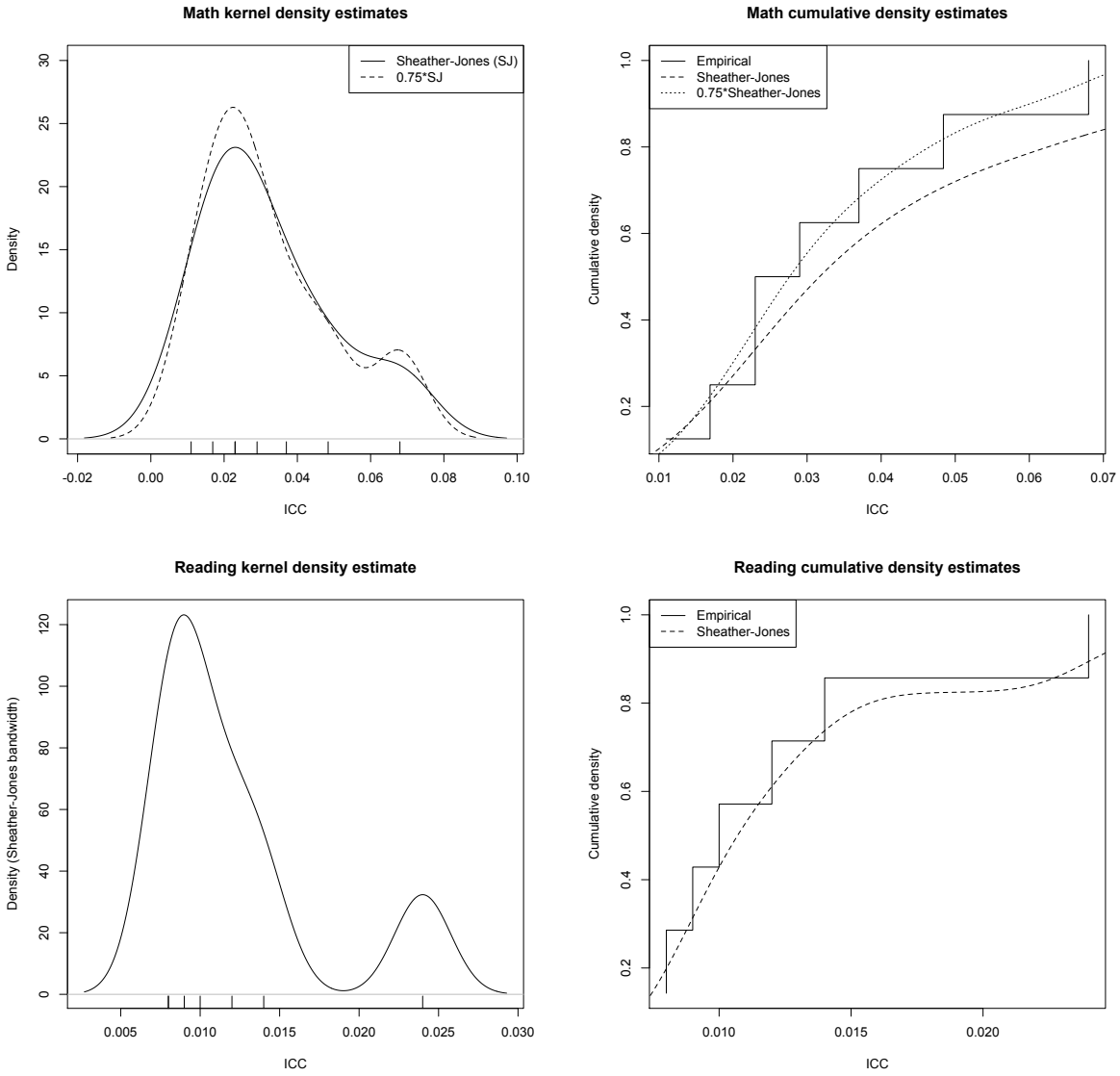

Figure 6. Kernel density estimates and comparisons of cumulative density estimates with empirical cumulative densities. The small vertical lines on the x-axes of the kernel density estimate graphs represent the original published estimates.

The sampling variability included in empirical densities raises concerns about overfitting to estimates from small studies in the tails of the distribution (Silverman, 1986). To represent the increased precision of estimates derived from larger samples, Rotondi and Donner (2009) generated kernel density estimates from published ICC estimates weighted by the inverses of

their estimated variances. Figure 7 compares weighted and unweighted estimates, using the estimated variance formula used by Hedges and Hedberg (2007), originally from Donner and Koval (1982):

$$Var(\hat{\rho}) = \frac{2(1-\hat{\rho})^2[1+(n-1)\hat{\rho}]^2}{\frac{N}{k}\left(\frac{N}{k}-1\right)k} \tag{19}$$

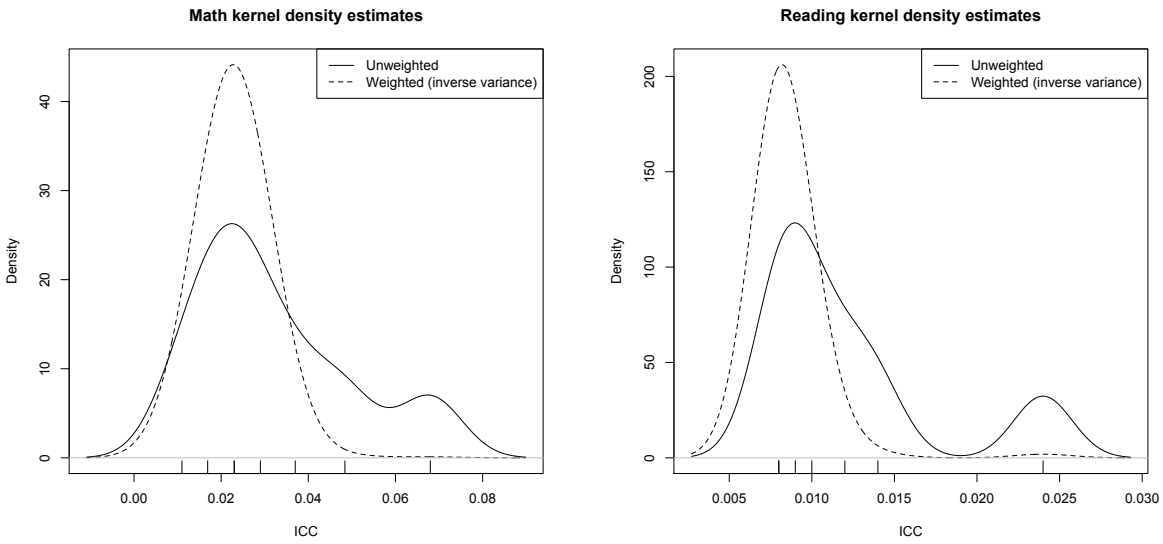

Figure 7. Kernel density estimates, unweighted and weighted by inverse of estimated variance. The small vertical lines on the x-axes represent the original published estimates.

For the math estimates, the Koedel and Betts (2011) and Jacob and Lefgren (2008) estimates are downweighted due to their relatively small sample sizes (389 and 201 teachers respectively). The Kane and Staiger (2008) reading estimate is also downweighted.

Comparing Fully Bayesian and Empirical Bayes Design Priors

For the empirical Bayes design priors, weighting the estimates by the inverses of their estimated variances partially accounts for sampling error by downweighting estimates from small studies with larger expected errors. However, the weighted densities still include some

sampling error, while the fully Bayesian distributions only represent uncertainty about ρ itself. Figure 7 shows that weighting the estimates has a large impact on the estimated distributions, but sampling error still makes the empirical Bayes distributions wider than the fully Bayesian distributions in Figure 4.

Rotondi and Donner (2009) raised a concern that common distributional assumptions, such as those used for the fully Bayesian design priors, may not describe published estimates well. A Bayesian approach to assessing model fit involves comparing the observed published estimates with simulated replicated datasets from the fitted model. The results suggest that the Fisher likelihood is a better fit than Swiger for both the math and reading estimates, and appears to fit the reading estimates well. Although both likelihoods show lack of fit to the Jacob and Lefgren (2008) math estimate, this is a small study (201 teachers) and so is subject to more sampling variability. It also uses student-level covariates, which Chapter 4 showed tend to be associated with larger between-teacher ICC estimates. Since the Fisher likelihood is a good fit to most of the estimates, this distribution will be the fully Bayesian prior used in the next stage of analysis.

The rationale for developing these design priors was the need to account for uncertainty about between-teacher ICC when performing prospective analysis for VAM. The next chapter will describe an approach to prospective power analysis using the design priors developed in this chapter.

CHAPTER 6: VAM Power Analysis with Design Priors

The ultimate goal of this dissertation study is probabilistic description of VAM power for teacher comparisons, using design priors for between-teacher ICC derived in Chapter 5 from published estimates. The approach to power calculation is hybrid: Bayesian design priors introduce estimation uncertainty into a classical prospective analysis. The intent is to evaluate the likelihood that a new VAM study, say for a previously unstudied district, would have reasonable power to identify teachers whose student scores differed from the district average by a stated threshold.

The disappointing power results of Schochet and Chiang (2013) for a single year of data have been reviewed in Chapter 2. The current study extends Schochet and Chiang's simulation by treating the between-teacher ICC as a random variable instead of a known, fixed value. This means that the ICC that will be estimated from a new VAM study is conceptualized as falling somewhere in a range of values, with varying likelihoods of specific values. If the ICC is on the larger end of this range, comparisons between teachers will have greater power. On the other hand, there is also some probability of smaller ICC values and lower power. The question for the prospective analysis was the range of power values associated with the range of plausible values of the ICC.

The prospective analysis can be briefly described as follows. Many possible values of the between-teacher ICC estimate $\hat{\rho}_T$ were drawn from a design prior distribution. These were used in equation 21 below to generate corresponding power values. These represented distributions for power derived from the design prior distributions of $\hat{\rho}_T$ – in other words, probabilities of achieving various levels of power, based on an assumed distribution of $\hat{\rho}_T$. This chapter provides details of the design and results of the power simulations.

Design of Simulations

Simulating Between-Teacher ICC Values from Design Priors

As discussed in Chapter 5, two sources of uncertainty about the between-teacher ICC estimates $\hat{\rho}_T$ are relevant. The first is that the underlying ICC is conceptualized as having a probability distribution. The second is that the ICC is estimated with sampling variability. Fully Bayesian design priors explicitly separate these two sources of uncertainty, while the empirical Bayes (kernel density) design priors combine them. Therefore, the two types of priors required two different methods of randomly drawing $\hat{\rho}_T$ values for the simulations.

Simulation from fully Bayesian design priors. Analysis of simulated data sets in Chapter 5 indicates that the Fisher distribution is a better fit to the published estimates than the Swiger distribution, so Fisher design priors were used for the simulations. Randomly drawing $\hat{\rho}_T$ values from these priors required several steps.

1. A random ICC value was drawn from the set of simulated ρ_T values from the MCMC process. This was then transformed as $g(\rho_T) = \frac{1}{2}\log\left[\frac{1+(N/k-1)\rho_T}{1-\rho_T}\right]$, where N = s*m*n (total number of students) and k = s*m (total number of teachers). See equation 16.
2. The (transformed) variance was calculated as $Var(\hat{\rho}_T) \sim \frac{1}{2}\{(k-1)^{-1} + (N-k)^{-1}\}$.
3. A simulated estimate was randomly drawn as $g(\hat{\rho}_T) \sim N\big(g(\rho), Var(\hat{\rho}_T)\big)$ and back-transformed to the original scale to obtain $\hat{\rho}_T$.

Simulation from Empirical Bayes design priors. Unlike the Fisher priors, kernel density design priors combine both sources of uncertainty. The unweighted kernel densities assign all studies equal weight regardless of sample size. Weighted densities, on the other hand, give larger studies larger weights and so partially adjust for sampling error. Values of $\hat{\rho}_T$ for the

simulations were drawn directly from the kernel densities; the probability of drawing a particular value of $\hat{\rho}_T$ was the kernel density at that value.

Model and Power Calculation

The power analysis assumed a three-level (students, teachers, and schools) model. The outcome variable g_{ijk} is the gain score for student i of teacher j in school k:

Level 1: Students i = 1, …, n $\quad g_{ijk} = \tau_{jk} + \varepsilon_{ijk}, \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ (20)

Level 2: Teachers j = 1, …, m $\quad \tau_{jk} = \eta_k + \theta_{jk},\ \theta_{jk} \sim N\left(0, \sigma_\theta^2\right)$

Level 3: Schools k = 1, …, s $\quad \eta_k = \delta + \psi_k,\ \psi_k \sim N\left(0, \sigma_\psi^2\right)$

Here δ represents the expected district-wide student score gain.

For the power analysis, the null hypothesis of interest is $H_0: \tau_{jk} - \bar{\bar{\tau}}_{..} = 0$, where $\bar{\bar{\tau}}_{..}$ is the average performance in the district. Chapter 2 introduced the equation for power to detect a teacher whose effect differs from the average by a threshold T, assuming a balanced design (same number of teachers m in each school, and same number of students n for each teacher). The power β was calculated as follows, where Φ[.] is the normal distribution, T is the threshold value, and model variances $\sigma_\psi^2, \sigma_\theta^2, \sigma_\varepsilon^2$ are as shown in equation 1:

$$\beta = \Phi\left[\frac{|T|\sigma}{2\sqrt{V_{est}}}\right] \tag{21}$$

$$\sigma = \sqrt{\sigma_\psi^2 + \sigma_\theta^2 + \sigma_\varepsilon^2}$$

Since variance estimates in the literature are standardized to unit total variance, variances are equal to ICCs.

Simulation Parameters

For comparison with Schochet and Chiang (2013), this study used their parameter values, including sample sizes n = 21 students per class, m = 10 teachers per school, s = 5 or 30 schools

per district, and between-school ICC of 0.005. This between-school ICC is the mean of values from recent studies.[7]

Chapter 3 described the subject-specific thresholds calculated from the different average annual growth estimates for math and reading. The math thresholds were 0.12, 0.24, and 0.36 standard deviations. The thresholds for reading were 0.08, 0.16, and 0.24 standard deviations.

Results of Simulations

Between-Teacher ICC Values Simulated from Design Priors

Figure 8 compares boxplots of $\hat{\rho}_T$ values simulated from the fully Bayesian Fisher distributions and from the kernel densities (10,000 simulations from each density). For the Fisher distributions, the underlying ρ_T distribution was constrained to [0, 1] using a Uniform[0, 1] prior. However, sampling variability resulted in random $\hat{\rho}_T$ values less than 0, especially for s = 5 schools (k = 50 teachers, N = 1050 students). For the power calculations in the next section, simulated $\hat{\rho}_T$ values less than zero were set to zero.

The 2.5th - 97.5th percentile range of the simulated $\hat{\rho}_T$ values represents a 95% "credible" interval, a Bayesian analog to a classical 95% confidence interval. For the math $\hat{\rho}_T$ values in Figure 8, all the 95% credible intervals exclude 0, while for the reading $\hat{\rho}_T$ values, the Fisher credible intervals include 0. These are consistent with the reading confidence interval from classical meta-analysis (Chapter 4), which also includes 0.

The Fisher credible intervals for s = 30 schools (300 teachers) are most similar to the classical meta-analysis confidence interval, which seems intuitively reasonable because this is

[7] Schochet and Chiang adjusted their school, classroom, and teacher ICC estimates, representing proportions of variance of posttest scores, for use in the model representing score gains. They divided the ICC estimates by an estimated ratio of the variance in score gains to the variance of posttest scores, which they calculated from individual-level data available to them. This adjustment seems to incorrectly inflate the school and teacher ICCs at the expense of the residual variance, and in any case has little impact on the ultimate power calculations. This dissertation study uses unadjusted school and teacher ICC estimates.

roughly the lower bound of the number of teachers per study for the published estimates. Meanwhile, the intervals for s = 5 schools show greater variability from the smaller sample. The impact of weighting the kernel densities appears as narrower interquartile ranges for the weighted densities, but the 95% credible intervals are unaffected.

For reading, the kernel density credible intervals are approximately the same width as the Fisher credible interval for s = 30 schools, although they are shifted slightly higher. However, for math, the kernel density credible intervals are wider than the Fisher s = 30 credible interval. The analysis using simulated data sets in Chapter 5 suggests that the Fisher likelihood is a reasonable fit to the math estimates. The kernel density credible intervals are wider due to the mass in the right tail from Jacob and Lefgren (2008) (see Figure 6).

The great majority of the simulated $\hat{\rho}_T$ values were larger than 0.005, the between-school ICC used in the simulations. This is consistent with a consensus in the literature that between-teacher variation represents a larger proportion of overall variation than between-school variation (e.g., Hanushek et al., 2005; Kane & Staiger, 2008; Nye et al., 2004).

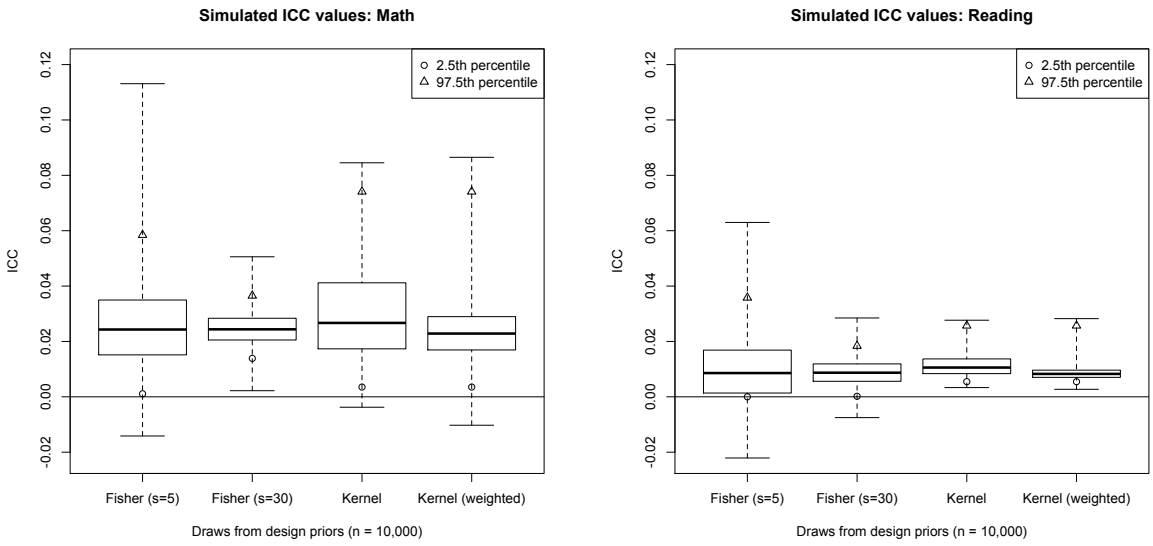

Figure 8. Summary of $\hat{\rho}_T$ values drawn from design prior distributions.

Power Distributions

The 10,000 simulated $\hat{\rho}_T$ values described in the previous section were used in equation 21 to calculate 10,000 corresponding power values. Figure 9 shows densities of the resulting power values for s = 5 and 30 schools and the smallest threshold T (0.12 standard deviations for math, 0.08 for reading). These densities can be interpreted as smooth curves drawn over histograms of the simulated power values, so that results from different types of estimates can be compared on the same graph.

Figure 9 shows that the variation in the $\hat{\rho}_T$ values has little effect on power. For math, power varies in a narrow range of 60.9% - 61.5%, and for reading in an even narrower range of 57.3%-57.5%. In general, more power values near the top of this range are observed with $\hat{\rho}_T$ values from the unweighted kernel design priors, because these priors have mass at higher ICC values from estimates in the right tail.

Table 5 presents percentiles of simulated power values for low, middle, and high thresholds from unweighted kernel design priors. Since these priors generate the greatest number of power values near the top of the range, this puts the power distribution in the best possible light. The percentiles in Table 5 can be interpreted as probabilities that power achieved will be less than the percentile value. For example, for the low math threshold, the 90th percentile for power is 61.3%; this indicates a 90% probability that power will not exceed 61.3%. Table 5 shows that differences in math and reading ICC distributions have some consequences for power. This can be seen especially in the high thresholds, treating the commonly used cutoff value of 80% power as the minimum acceptable. For the high math threshold, the 10th percentile is 80%, so the probability of power below 80% is low. On the other hand, since the 90th percentile for the high reading threshold is only 71.3%, it is very likely that power will fall below 80%.

In summary, ICC variation within subject (math or reading) in the range supported by the published estimates has little effect on power to detect differences between teachers. Using the three-level HLM, 80% power is highly unlikely, except using the largest threshold value for math.

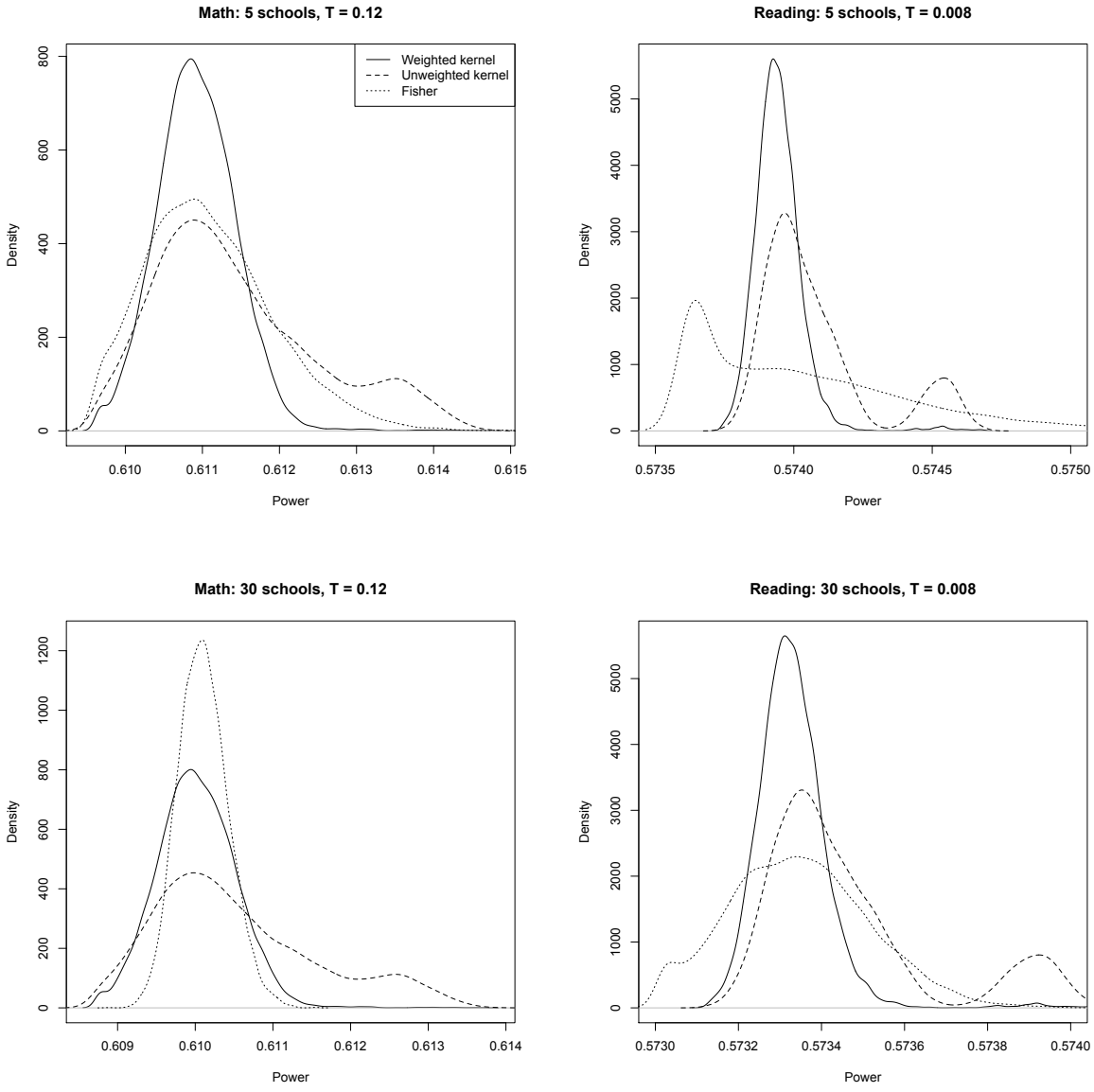

Figure 9. Densities of simulated power values. These can be interpreted as smooth curves drawn over histograms of the simulated power values, so that multiple results can be compared on the same graph.

Table 5

Percentiles of Simulated Power Values from Unweighted Kernel Design Priors

Subject	**Threshold (sd)**	**10th percentile**	**50th percentile**	**90th percentile**
Math	Low (0.12)	0.610	0.611	0.613
	Medium (0.24)	0.712	0.714	0.717
	High (0.36)	0.799	0.802	0.806
Reading	Low (0.08)	0.574	0.574	0.574
	Medium (0.16)	0.645	0.646	0.646
	High (0.24)	0.712	0.712	0.713

CHAPTER 7: Discussion and Conclusions

This dissertation has described an innovative application of hybrid Bayesian and classical prospective analysis techniques to power analysis for VAM teacher comparisons. This final chapter summarizes major findings of the study and implications for research and policy uses of VAM. It also discusses the study's limitations and possible directions for future work, "had we but world enough, and time," as the poet Andrew Marvell wrote.

Review of Published Between-Teacher ICC Estimates

Why are Math Estimates Larger?

The VAM literature has previously noted that between-teacher variance estimates from math test scores tend to be larger than estimates from reading test scores (Hanushek & Rivkin, 2010; Nye et al., 2004). The current study confirmed this in a review of 31 estimates from eight VAM studies published in the last ten years. Besides being generally larger, the math estimates have a wider spread. All studies but one reported both math and reading estimates using the same model specifications and similar sample sizes, so these factors do not explain the differences between subjects. Chapter 4 provided confidence intervals for the variance estimates from classical meta-analysis, and Chapter 6 reported credible intervals from Bayesian methods. The meta-analysis generates a 95% confidence interval for reading ICC that includes zero, as do the Bayesian Fisher credible intervals. A zero ICC can be interpreted as no correlation of reading scores between students of the same teacher. The 95% confidence and credible intervals for math ICC, on the other hand, exclude zero. Chapter 6 showed that the differences between math and reading estimates affect predictions of power to detect a teacher who differs from average, a finding further discussed later in the chapter.

The VAM literature has been almost silent on the subject of the larger math estimates. Nye et al. (2004) attributed the larger between-teacher variance estimates for math to systematic differences in learning and/or instruction: "This may be because mathematics is mostly learned in school and thus may be more directly influenced by teachers, or that there is more variation in how (or how well or how much) teachers teach mathematics. Reading, on the other hand, is more likely to be learned (in part) outside of school and thus the influence of school and teacher on reading is smaller, or there is less variation in how (or how well or how much) reading is taught in school" (p. 247). If this interpretation is valid, the wider spread of the math estimates implies that some states/districts show more variation in math teacher effectiveness than others. In other words, in some areas, the difference between the most and least effective math teachers is large, while other areas are more homogeneous; meanwhile, differences between areas are not as pronounced for reading teachers.

Other potential explanations for the larger math estimates involve differences in the instruments used to measure math and reading achievement. As Papay (2011) wrote, "tests differ in their content, scaling, samples of students who take them, and timing, all of which can introduce inconsistency into estimates of true teacher effectiveness" (p. 183). One example of a test characteristic that could directly impact teacher effect estimates is *instructional sensitivity.* This term is used in the measurement literature for an assessment's ability to detect student achievement differences associated with instructional opportunities and teacher effectiveness. Measurement researchers have repeatedly raised concerns about the instructional sensitivity of large-scale standardized tests (e.g., Popham, 2007; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). The small between-teacher variance estimates from reading tests may be a sign that these tests are somewhat insensitive to instruction.

Although comparisons of results from different test instruments have begun to appear in the VAM literature (Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez, 2007; Papay, 2011), the role of instructional sensitivity has yet to be directly addressed. Papay (2011) compared VAM teacher rankings from three different reading tests given the same year in the same district, and found only moderate rank correlations, showing a clear impact of the choice of test instrument on estimates. Papay unfortunately did not report between-teacher variance estimates from the three instruments. Also, all three tests – an unidentified state test, the Stanford Achievement Test, and the Scholastic Reading Inventory – would be considered "distal" by Ruiz-Primo et al. (2002). That is, they assess broad state or national reading standards rather than specific curricula. Research focusing on implications of instructional sensitivity might compare between-teacher variance estimates from instruments of varying sensitivity. A study could, for example, compare results from distal reading tests with results from district tests tailored to specific reading curricula, which would be expected to show greater sensitivity to instruction.

Model Specifications and Classical Meta-Analysis

The literature has not previously compared between-teacher variance estimates from studies using different VAM specifications, except within individual studies. Chapter 4 included a first attempt to identify systematic differences in estimates calculated from different models. The specifications compared included fixed effects, random effects, and covariate adjustment at the school level, and fixed effects and covariate adjustment at the student level. The single study using school random effects (Nye et al., 2004) obtained between-teacher variance estimates much larger than those of other studies. Estimates also tended to be larger in models with student covariate adjustment than in models with student fixed effects.

Quantifying the expected effects of specification differences would facilitate more straightforward comparisons of between-teacher ICC estimates across VAM studies. A limitation of the classical meta-analysis in Chapter 4 is that it does not take advantage of model specifications as covariate information; it could be extended using meta-regression techniques (Hartung et al., 2008). A meta-regression could estimate, for example, the increase in between-teacher variance expected from specifying student covariates instead of student fixed effects. Meta-regression would be enhanced by the greater flexibility in specification offered by Bayesian estimation (Spiegelhalter et al., 2004).

Bayesian Design Priors

This dissertation introduces to the educational research literature the use of fully Bayesian design priors to represent empirical evidence about ICC values. The fully Bayesian priors were compared to kernel density estimates like those of Rotondi and Donner (2009), representing empirical Bayes priors. The findings in this section pertain to the methodological question of how well the observed ICC values are described by the priors used in the study.

Rotondi and Donner (2009) explored kernel density priors because they were concerned that common distributional assumptions, such as those used for the fully Bayesian design priors, may not describe published estimates well. A special concern for this study was whether the Swiger and Fisher likelihoods, developed for randomized studies (Ukomunne, 2002), made sense for VAM observational studies. A Bayesian fit assessment in Chapter 5 compared the published estimates with simulated datasets generated from the likelihoods. The results suggest that the Fisher likelihood is a better fit than Swiger for both the math and reading estimates, and appears to fit the reading estimates well. Both likelihoods show lack of fit to the Jacob and Lefgren (2008) math estimate. On the other hand, the unweighted kernel density fit for math faithfully

reproduces this estimate as a peak in the right tail (Figure 6). A caveat regarding kernel density fits is that they are known to overfit to observations in the tails of the distribution (Silverman, 1986). In the inverse variance-weighted kernel density fit, the peak disappears, because the Jacob and Lefgren study is small (201 teachers) and thus subject to relatively high sampling variability. This suggests that the unweighted fit does in fact overfit to Jacob and Lefgren (2008). Overall, the evidence is in favor of the Fisher likelihood as a good description of the estimates.

A fully Bayesian prior like the Fisher prior has methodological advantages over kernel density priors. For example, it is straightforward to incorporate new information (in this case, newly published estimates) using Bayesian updating. The design prior simply becomes the new initial prior, and the same distributional form for the likelihood is used to generate a new design prior that incorporates the new estimates. Empirical densities, on the other hand, cannot be updated in the same way; incorporating new estimates would require a new fit. Another advantage may be especially important in research focused on specific states or districts. A fully Bayesian prior for a particular area can be generated from any relevant estimates, even if only one or two estimates are available. This would not be feasible with kernel densities; at least four estimates are recommended for reasonable density approximation in R (Rotondi & Donner, 2009). A fully Bayesian approach also offers greater flexibility for incorporating information about estimates with varying levels of relevance to the area of interest (Turner, Thompson, and Spiegelhalter, 2005).

Research interest in VAM is intense, so new estimates of between-teacher ICC from additional geographical areas can be expected to appear in the literature. Design priors can be interpreted as predicting the probabilities of observing different values of ICC. Future research could examine whether newly published estimates represent probable values under the Fisher

prior. If predictions made by the prior are borne out by future estimates, this would further support their use in future analyses.

Power for VAM Teacher Comparisons

This dissertation study has described hybrid prospective analysis of VAM, combining Bayesian design priors with a classical power analysis. Chapter 6 shows that the empirically supported variation in between-teacher ICC within subject (math or reading) has little effect on power to detect teachers who differ from average. Table 6 summarizes power results for the math and reading estimates. This study's key educational policy implication is that it strengthens the evidence that the small contribution of between-teacher variance to total variance limits the utility of VAM for teacher comparisons. Acceptable (at least 80%) power can be achieved for one year of data only for math scores, and only for a large threshold difference of a teacher from the average: 36% of average annual gain.

The analysis here used a three-level hierarchical model with students nested within teachers nested within schools. Schochet and Chiang (2013) used a four-level model, with levels for students, *classrooms,* teachers, and schools. This model included a between-classroom variance component estimated from classrooms taught by the same teacher across multiple years. Schochet and Chiang used the four-level model in order to calculate the power improvement achieved by combining several years of data into a single estimate for each teacher.

Table 6

Comparison of Power Values (Single Year of Data)

Threshold (standard deviations)	**Current study – Math (50^{th} percentile)**	**Current study – Reading (50^{th} percentile)**	**Schochet & Chiang (2013)** $\hat{\rho}_T$= 0.035
Low (math = 0.12; reading = 0.08; Schochet & Chiang = 0.1)	0.61	0.57	0.57
Medium (math = 0.24; reading = 0.16; Schochet & Chiang = 0.2)	0.71	0.65	0.64
High (math = 0.36; reading = 0.24; Schochet & Chiang = 0.3)	0.80	0.71	0.71

Variation in between-teacher ICC values affects power by influencing the residual variance σ_ε^2, which impacts the variance of the teacher effect estimator. For the three-level model, the variance of the estimator is $V_{OLS} = \frac{\sigma_\varepsilon^2}{n}\left(\frac{sm-1}{sm}\right)$, as discussed in Chapter 3. For the four-level model, it is:

$$V_{OLS} = \left(\frac{\sigma_\omega^2}{c} + \frac{\sigma_\varepsilon^2}{cn}\right)\left(\frac{sm-1}{sm}\right) \quad (22)$$

where σ_ω^2 is the between-classroom variance and c is the number of years of data. Between-teacher ICC variation in the empirically supported range has minimal effect on σ_ε^2. Thus this variation does not strongly affect power for the three-level model, nor would it greatly impact multiple-year power calculations assuming a four-level model.

Schochet and Chiang's analysis indicated that acceptable power for teacher comparisons can be attained only if multiple years of data for a teacher are combined into a single estimate. Although this is common accountability practice (National Research Council and National Academy of Education, 2010), it assumes that the measurement properties of the test instruments do not change over time. This may not be the case if, for example, performance patterns change as students and teachers become more familiar with the test format (Linn, Graue, & Sanders, 1990). Also, different forms of a test instrument from different years must be adequately equated. Combining data from multiple years into a single estimate also obscures changes in teacher effectiveness over time, resulting for example from experience (Plecki, Elfers, & Nakamura, 2012). Even if these caveats are reasonable, the results of the current study provide additional evidence that VAM is unsuitable for evaluating new teachers or other teachers for whom multiple consecutive years of data are unavailable.

Conclusions and Significance

This dissertation has demonstrated the use of Bayesian design priors to represent uncertainty about between-teacher ICC in a hybrid VAM prospective analysis with a classical power calculation. Several important findings have been summarized in this chapter. The review of published estimates confirmed unexplained differences in ICC distribution between math and reading estimates. Evidence has been presented that the Fisher likelihood is a reasonable representation of the available estimates for use as a Bayesian design prior. Finally, it has been shown that the empirically supported within-subject variation in between-teacher ICC values has little effect on VAM's power for teacher comparisons. However, under the assumptions of the study, somewhat better power can be expected for math teacher comparisons than for reading teacher comparisons.

The utility of the techniques used in this dissertation extends beyond the power analysis presented here to other prospective analyses for VAM. This work has treated between-school and between-classroom ICCs as fixed, known quantities, when in fact they are also estimated with uncertainty. It would be a straightforward extension to determine how uncertainty of estimation of all ICCs would impact power for between-teacher comparisons. A related study could examine power for between-school comparisons. These extensions illustrate the flexibility of this dissertation's methods, whose applicability extends even beyond VAM to hierarchical analyses more generally.

REFERENCES

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. Economic Policy Institute Briefing Paper #278.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 103-116.

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The kernel approach with S-Plus illustrations.* New York: Oxford.

De Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*(1), 95-113.

Donner, A., & Koval, J. J. (1982). Design considerations in the estimation of intraclass correlation. *Annals of Human Genetics, 46*, 271-277.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review: Papers and Proceedings 100,* 267-271.

Hartung, J., Knapp, G., & Sinha, B. K. (2008). *Statistical meta-analysis with applications.* Hoboken, NJ: Wiley.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Jacob, B. A., & Lefgren, L. (2005). Principals as agents: Subjective performance measurement in education. National Bureau of Economic Research Working Paper 11463.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics 26*(1), 101-136.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review 27*, 615-631.

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. National Bureau of Economic Research Working Paper 14607.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy 6*(1), 18-42.

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. National Bureau of Economic Research Working Paper 14607.

Kruschke, J. K. (2011). *Bayesian data analysis: A tutorial with R and BUGS*. Amsterdam: Elsevier.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of claims that "everyone is above average". *Educational Measurement: Issues and Practice* (Fall), 5-14.

Loader, C. (1999). *Local regression and likelihood*. New York: Springer.

Lockwood, J. R., & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics, 1*, 223-252 (electronic). DOI: 10.1214/07-EJS057.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement 44*(1), 47-67.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book: A practical introduction to Bayesian analysis.* Boca Raton, FL: CRC Press.

Lunn, D., Spiegelhalter, D., Thomas, A. & Best, N. (2009) The BUGS project: Evolution, critique and future directions, *Statistics in Medicine 28*, 3049-3082.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 121-129.

National Research Council and National Academy of Education (2010). Getting value out of value-added: Report of a workshop. Washington, DC: The National Academies Press.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237-57.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal 48*(1), 163-193.

Plecki, M. L., Elfers, A. M., & Nakamura, Y. (2012). Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value-added measures. *Journal of Teacher Education 63*(5), 318-334.

Popham, W. J. (2007). Instructional sensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*(2), 146-150, 155.

R Foundation for Statistical Computing (2012). R: A Language and Environment for Statistical Computing. http://www.R-project.org.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review (AEA Papers and Proceedings), 94*(2), 247-252.

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

Rotondi, M. A., & Donner, A. (2009). Sample size estimation in cluster randomized educational trials: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 34*(2), 229-237.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369-393.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower-poverty schools. *Journal of Urban Economics, 72*, 104-122.

Schochet, P. Z., & Chiang, H. S. (2010). Error rates in measuring teacher and school performance based on student test score gains. NCEE 2010-4004.

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, *38*(2), 142-171.

Sheather, S. J. (2004). Density estimation. *Statistical Science, 19*(4), 588-597.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* London; New York: Chapman and Hall.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health care evaluation*. Chichester; Hoboken NJ: Wiley.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., & Lunn, D. (1994, 2002). BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England. www.mrc-bsu.cam.ac.uk/bugs/

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. Journal of Statistical Software, 12(3), 1-16.

Turner, R. M., Prevost, A. T., & Thompson, S. G. (2004). Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine, 23*(8), 1195-1214.

Turner, R. M., Thompson, S. G., & Spiegelhalter, D. J. (2005). Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials, 2,* 108-118.

Ukoumunne, O. C. (2002). A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine, 21*, 3757-3774.

APPENDIX

R and OpenBUGS Code

Appendix: R and OpenBUGS code

Chapter 4

```
# classical meta-analysis - Hartung, Knapp, Sinha (2008) p. 37

ve = read.csv("est4bayes1.csv", stringsAsFactors=FALSE)
ve.m = ve[ve$Subject == "Math",]
ve.r = ve[ve$Subject == "Reading",]

ve.s = ve

# var of the correlation estimates
var.r = (1 - ve.s$var_est^2)^2/(ve.s$k - 1)

sum1 = sum(ve.s$var_est/var.r)
sum2 = sum(1/var.r)
sum3 = sum(ve.s$var_est^2/var.r)

# combined estimate
theta.hat = sum1/sum2

# 95% CI
var.theta.hat = 1/sum2
c(theta.hat - 1.96*sqrt(var.theta.hat), theta.hat +
 1.96*sqrt(var.theta.hat))

# homogeneity test
chi2.h = sum3 - sum1^2/sum2
qchisq(p=.05, df=dim(ve.s)[1]-1, lower.tail = FALSE)
```

Chapter 5

Fully Bayesian design priors
(Math estimates shown; reading used same code with different data)

```
library(R2OpenBUGS)
ve = read.csv("est4bayes1.csv", stringsAsFactors=FALSE)
ve.m = ve[ve$Subject == "Math",]
```

FISHER

Model file for OpenBUGS

```
model {

  for (m in 1:r) {

    # Fisher transformation
    g.rho[m] <- 0.5*log( ( 1 + (N[m]/(k[m]-1))*rho )/(1 - rho) )



    # Fisher formula for the variance
    var.rho.hat[m] <- 0.5*( 1/(k[m]-1) + 1/(N[m]-k[m]) )

    tau.rho.hat[m] <- pow(var.rho.hat[m], -1)

    rho.hat[m] ~ dnorm(g.rho[m], tau.rho.hat[m])

  }

  # PRIORS

  # Turner 2004
  rho ~ dunif(0, 1)

}
```

R code

```
# Fisher transformation
rho.hat = 0.5*log( (1 + (ve.m$N/ve.m$k - 1)*ve.m$var_est) / (1 -
 ve.m$var_est) )
r = length(ve.m$var_est)
N = ve.m$N
k = ve.m$k
data = list("rho.hat", "r", "N", "k")

# parameters to be monitored:
parameters = c("rho")

inits = function()
{
  list( "rho"=runif(1) )
}

# Call to OpenBUGS
samples = bugs(data, inits, parameters,
 model.file ="Fisher_FE.txt",
 n.chains=3, n.iter=1000, n.burnin=100, n.thin=1, DIC=TRUE,
```

```
 codaPkg=FALSE, debug=TRUE, working.directory=getwd())

# Convergence
gelman.out <- gelman.diag(samples, autoburnin=FALSE)
gelman.out$psrf
gelman.out$mpsrf
```

SWIGER

Model file for OpenBUGS

```
model {

  for (m in 1:r) {

    # Swiger formula for the variance
    var.rho.hat[m] <- ( 2*(N[m]-1)*pow(1-rho, 2)*pow(1+(N[m]/k[m]-
1)*rho, 2) )/( pow(N[m]/k[m], 2)*(N[m]-k[m])*(k[m]-1) )

    tau.rho.hat[m] <- pow(var.rho.hat[m], -1)

    rho.hat[m] ~ dnorm(rho, tau.rho.hat[m])

  }

  # PRIORS

  # Turner 2004
  rho ~ dunif(0, 1)

}
```

R code

```
rho.hat = ve.m$var_est
r = length(ve.m$var_est)
N = ve.m$N
k = ve.m$k
data = list("rho.hat", "r", "N", "k")

# parameters to be monitored:
parameters = c("rho")

inits = function()
{
  list( "rho"=runif(1) )
}

# Call to OpenBUGS
samples = bugs(data, inits, parameters,
```

```
 model.file ="SwigerFE.txt",
 n.chains=3, n.iter=1000, n.burnin=100, n.thin=1, DIC=TRUE,
 codaPkg=FALSE, debug=TRUE, working.directory=getwd())

# Convergence
gelman.out <- gelman.diag(samples, autoburnin=FALSE)
gelman.out$psrf
gelman.out$mpsrf
```

Fit analysis using simulated data sets (see Figure 5)

```
fake.swiger <- function(data, samples, n.sims)
{
  est.mat = NULL
  dens.rho = density(samples$sims.list$rho)
  rho.hat = data$var_est; N = data$N; k = data$k

  for (i in 1:n.sims) {
    # draw rho value
    rho = sample(dens.rho$x, size=1, prob=dens.rho$y)

    # draw vector of estimates
    est.vec = vector(mode = "numeric", length = length(rho.hat))
    for (m in 1:length(rho.hat)) {
 est.var = ( 2*(N[m]-1)*(1-rho)^2*(1+(N[m]/k[m]-1)*rho)^2 )/(
 (N[m]/k[m])^2*(N[m]-k[m])*(k[m]-1) )
 est.vec[m] = rnorm(n=1, mean=rho, sd=sqrt(est.var))
    }
    est.mat = rbind(est.mat, est.vec)
  }

  tmp = apply(est.mat, 1, "summary")
  meta = apply(tmp, 1, "summary")
  rownames(meta) = c("min of", "1Q of", "median of", "mean of", "3Q
 of", "max of")
  meta
}

fake.fisher <- function(data, samples, n.sims)
{
  est.mat = NULL
  dens.rho = density(samples$sims.list$rho)
  rho.hat = data$var_est; N = data$N; k = data$k

  for (i in 1:n.sims) {
    # draw rho value
    rho = sample(dens.rho$x, size=1, prob=dens.rho$y)

    # draw vector of estimates
    est.vec = vector(mode = "numeric", length = length(rho.hat))
    for (m in 1:length(rho.hat)) {
```

```
 g.rho = 0.5*log( (1 + (N[m]/(k[m]-1))*rho )/(1 - rho) )
 est.var = 0.5*( 1/(k[m]-1) + 1/(N[m]-k[m]) )
 est.vec[m] = rnorm(n=1, mean=g.rho, sd=sqrt(est.var))
    }
    est.mat = rbind(est.mat, est.vec)
  }

  tmp = apply(est.mat, 1, "summary")
  meta = apply(tmp, 1, "summary")
  rownames(meta) = c("min of", "1Q of", "median of", "mean of", "3Q
 of", "max of")
  meta
}
```

Gaussian kernel densities

```
ve = read.csv("est4bayes1.csv", stringsAsFactors=FALSE)
ve.m = ve$var_est[ve$Subject == "Math"]
ve.r = ve$var_est[ve$Subject == "Reading"]

# MATH

> bw.SJ(ve.m)
[1] 0.01

SJ.m <- density(ve.m, bw="SJ")
SJ75.m <- density(ve.m, bw="SJ", adjust=0.75)

# READING

> bw.SJ(ve.r)
[1] 0.002

SJ.r <- density(ve.r, bw="SJ")

# WEIGHTED by inverse of variance est

ve = read.csv("est4bayes1.csv", stringsAsFactors=FALSE)

# MATH
ve.m = ve[ve$Subject == "Math",]

# var of the corr est (Donner & Koval 1982 -> Hedges & Hedberg)
var.dk = 2*(1 - ve.m$var_est)^2*( 1+(ve.m$N/ve.m$k - 1)*ve.m$var_est
 )^2/( (ve.m$N/ve.m$k)*(ve.m$N/ve.m$k - 1)*(ve.m$k - 1) )


var.r = var.dk
inv.r = 1/var.r
inv.sum1 = inv.r/sum(inv.r)
SJ75w.m <- density(ve.m$var_est, bw="SJ", adjust=0.75,
```

```
 weights=inv.sum1)


# READING
ve.r = ve[ve$Subject == "Reading",]

# var of the corr est (Donner & Koval 1982 -> Hedges & Hedberg)
var.dk = 2*(1 - ve.r$var_est)^2*( 1+(ve.r$N/ve.r$k - 1)*ve.r$var_est
 )^2/( (ve.r$N/ve.r$k)*(ve.r$N/ve.r$k - 1)*(ve.r$k - 1) )

var.r = var.dk
inv.r = 1/var.r
inv.sum1 = inv.r/sum(inv.r)
SJw.r <- density(ve.r$var_est, bw="SJ", weights=inv.sum1)
```

Chapter 6

Simulate ICCs from Fisher design prior

```
fisher.sims <- function(samples, n.sims, s, m, n)
{
  # sample sizes
  k = s*m
  N = s*m*n

  est.vec = vector(mode="double", length=n.sims)
  rho = samples$BUGSoutput$sims.list$rho

  for (i in 1:n.sims) {
    # draw rho value
    rho = sample(rho, size=1)

    # draw estimate

    g.rho = 0.5*log( (1 + (N/(k-1))*rho )/(1 - rho) )
    est.var = 0.5*( 1/(k-1) + 1/(N-k) )
    rho.hat.t = rnorm(1, mean=g.rho, sd=sqrt(est.var))
    # back-transform
    est.vec[i] = -(1 - exp(2*rho.hat.t))/(N/k - 1 +
    exp(2*rho.hat.t))
  }

  est.vec

}

rho.m.f5 = fisher.sims(samples.m, n.sims=10000, s=5, m=10, n=21)
rho.m.f30 = fisher.sims(samples.m, n.sims=10000, s=30, m=10, n=21)
rho.r.f5 = fisher.sims(samples.r, n.sims=10000, s=5, m=10, n=21)
rho.r.f30 = fisher.sims(samples.r, n.sims=10000, s=30, m=10, n=21)
```

Simulate ICCs from kernel density design priors

```
# Unweighted
rho.m.uwk = sample(SJ75.m$x, size=10000, prob=SJ75.m$y, replace=TRUE)
rho.r.uwk = sample(SJ.r$x, size=10000, prob=SJ.r$y, replace=TRUE)
# Weighted
rho.m.wk = sample(SJ75w.m$x, size=10000, prob=SJ75w.m$y, replace=TRUE)
rho.r.wk = sample(SJw.r$x, size=10000, prob=SJw.r$y, replace=TRUE)
```

Calculate power

```
vam.power<- function(teacherICC, schoolICC, s, m, n, thresh)
{
 resid.var = 1 - (teacherICC + schoolICC)
 var.ols = (resid.var/n) * (s*m-1)/(s*m)
 pow = pnorm(thresh/(2*sqrt(var.ols)))
 pow
}

rho.m.uwk[rho.m.uwk < 0] = 0
rho.m.wk[rho.m.wk < 0] = 0
rho.m.f5[rho.m.f5 < 0] = 0
rho.m.f30[rho.m.f30 < 0] = 0
rho.r.uwk[rho.r.uwk < 0] = 0
rho.r.wk[rho.r.wk < 0] = 0
rho.r.f5[rho.r.f5 < 0] = 0
rho.r.f30[rho.r.f30 < 0] = 0

m.uwk = list()
m.uwk$s5.1 = vam.power(rho.m.uwk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.12)
m.uwk$s5.2 = vam.power(rho.m.uwk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.24)
m.uwk$s5.3 = vam.power(rho.m.uwk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.36)
m.uwk$s30.1 = vam.power(rho.m.uwk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.12)
m.uwk$s30.2 = vam.power(rho.m.uwk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.24)
m.uwk$s30.3 = vam.power(rho.m.uwk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.36)
m.wk = list()
m.wk$s5.1 = vam.power(rho.m.wk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.12)
m.wk$s5.2 = vam.power(rho.m.wk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.24)
m.wk$s5.3 = vam.power(rho.m.wk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.36)
m.wk$s30.1 = vam.power(rho.m.wk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.12)
m.wk$s30.2 = vam.power(rho.m.wk, schoolICC=0.005, s=30, m=10, n=21,
```

```
 thresh=0.24)
m.wk$s30.3 = vam.power(rho.m.wk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.36)
m.f = list()
m.f$s5.1 = vam.power(rho.m.f5, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.12)
m.f$s5.2 = vam.power(rho.m.f5, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.24)
m.f$s5.3 = vam.power(rho.m.f5, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.36)
m.f$s30.1 = vam.power(rho.m.f30, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.12)
m.f$s30.2 = vam.power(rho.m.f30, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.24)
m.f$s30.3 = vam.power(rho.m.f30, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.36)

r.uwk = list()
r.uwk$s5.1 = vam.power(rho.r.uwk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.08)
r.uwk$s5.2 = vam.power(rho.r.uwk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.16)
r.uwk$s5.3 = vam.power(rho.r.uwk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.24)
r.uwk$s30.1 = vam.power(rho.r.uwk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.08)
r.uwk$s30.2 = vam.power(rho.r.uwk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.16)
r.uwk$s30.3 = vam.power(rho.r.uwk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.24)
r.wk = list()
r.wk$s5.1 = vam.power(rho.r.wk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.08)
r.wk$s5.2 = vam.power(rho.r.wk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.16)
r.wk$s5.3 = vam.power(rho.r.wk, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.24)
r.wk$s30.1 = vam.power(rho.r.wk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.08)
r.wk$s30.2 = vam.power(rho.r.wk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.16)
r.wk$s30.3 = vam.power(rho.r.wk, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.24)
r.f = list()
r.f$s5.1 = vam.power(rho.r.f5, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.08)
r.f$s5.2 = vam.power(rho.r.f5, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.16)
r.f$s5.3 = vam.power(rho.r.f5, schoolICC=0.005, s=5, m=10, n=21,
 thresh=0.24)
r.f$s30.1 = vam.power(rho.r.f30, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.08)
```

```
r.f$s30.2 = vam.power(rho.r.f30, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.16)
r.f$s30.3 = vam.power(rho.r.f30, schoolICC=0.005, s=30, m=10, n=21,
 thresh=0.24)
```