INFORMATIVE PRIOR DISTRIBUTIONS IN MULTILEVEL/HIERARCHICAL LINEAR GROWTH MODELS: DEMONSTRATING THE USE OF BAYESIAN UPDATING FOR FIXED EFFECTS

by

ANDREW DANIEL SCHAPER

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of requirements
for the degree of
Doctor of Philosophy

June 2014

DISSERTATION APPROVAL PAGE

Student: Andrew Daniel Schaper

Title: Informative Prior Distributions in Multilevel/Hierarchical Linear Growth Models: Demonstrating the Use of Bayesian Updating for Fixed Effects

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Akihito Kamata	Chair
Gina Biancarosa	Core Member
Keith Zvoch	Core Member
Kent McIntosh	Institutional Representative

and

Kimberly Andrews Espy	Vice President for Research and Innovation; Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2014

DISSERTATION ABSTRACT

Andrew Daniel Schaper

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

June 2014

Title: Informative Prior Distributions in Multilevel/Hierarchical Linear Growth Models: Demonstrating the Use of Bayesian Updating for Fixed Effects

This study demonstrates a fully Bayesian approach to multilevel/hierarchical linear growth modeling using freely available software. Further, the study incorporates informative prior distributions for fixed effect estimates using an objective approach. The objective approach uses previous sample results to form prior distributions included in subsequent samples analyses, a process referred to as Bayesian updating. Further, a method for model checking is outlined based on fit indices including information criteria (i.e., Akaike information criterion, Bayesian information criterion, and deviance information criterion) and approximate Bayes factor calculations. For this demonstration, five distinct samples of schools in the process of implementing School-Wide Positive Behavior Interventions and Supports (SWPBIS) collected from 2008 to 2013 were used with the unit of analysis being the school. First, the within-year SWPBIS fidelity growth was modeled as a function of time measured in months from initial measurement occasion. Uninformative priors were used to estimate growth parameters for the 2008-09 sample, and both uninformative and informative priors based on previous years' samples were used to model data from the 2009-10, 2010-11, 2011-12, 2012-13 samples. Bayesian estimates were also compared to maximum likelihood estimates, and reliability

information is provided. Second, an additional three examples demonstrated how to include predictors into the growth model with demonstrations for: (a) the inclusion of one school-level predictor (years implementing) of SWPBIS fidelity growth, (b) several school-level predictors (relative socio-economic status, size, and geographic location), and (c) school and district predictors (sustainability factors hypothesized to be related to implementation processes) in a three-level growth model. Interestingly, Bayesian models estimated with informative prior distributions in all cases resulted in more optimal fit indices than models estimated with uninformative prior distributions.

CURRICULUM VITAE

NAME OF AUTHOR: Andrew Daniel Schaper

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene OR
San Francisco State University, San Francisco CA
The Colorado College, Colorado Springs CO

DEGREES AWARDED:

Doctor of Philosophy, Educational Methodology, Policy, and Leadership, 2014, University of Oregon
Bachelor of Arts, English, 2004, The Colorado College

AREAS OF SPECIAL INTEREST:

Quantitative Research Methodology
Growth Modeling
Bayesian Estimation

PROFESSIONAL EXPERIENCE:

Research assistant, Center on Teaching and Learning (CTL), University of Oregon, Eugene OR, 2011-2013
Research assistant, Center for Applied Second Language Studies (CASLS), University of Oregon, Eugene OR 2010-2011
Teacher: Middle and high school English language arts, San Francisco Unified School District (SFUSD), San Francisco CA, 2007-2010
Teacher: 5th and 6th grade humanities, The Town School for Boys, San Francisco CA, 2005-2007

GRANTS, AWARDS, AND HONORS:

Graduate Teaching Fellowship, Center on Teaching and Learning (CTL), University of Oregon, summer 2011 to fall 2013

Graduate Teaching Fellowship, Center for Applied Second Language Studies (CASLS), University of Oregon, fall 2010 to spring 2011

PUBLICATIONS:

Cummings, K.D., Biancarosa, G., Schaper, A., & Reed, D.K. (in press). Examiner error in curriculum-based measurement of oral reading. *Journal of School Psychology*

Reed, D.K., Cummings, K.D., Schaper, A., & Biancarosa, G. (in press). Assessment fidelity in reading intervention research: A synthesis of the literature. *Review of Educational Research*

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the professors in the College of Education at the University of Oregon. A special thank you to Dr. Akihito Kamata for providing guidance throughout my graduate studies as well as introducing me to the world of Bayesian statistics, Dr. Gina Biancarosa for mentoring me in classroom and professional settings, Dr. Keith Zvoch for his poignant feedback, and Dr. Kent McIntosh for allowing access to the data used for this project and providing on-going feedback.

Data for this project was supported from assistance from the federal government. Specifically, the research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120278 to University of Oregon. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

I would also like to thank those that allowed me to include previously published material in this dissertation including Drs. Karen Blase, Dean Fixsen, Andrew Gelman, Robert Horner, and Kent McIntosh.

Finally, I would like to thank my family and friends for their ongoing support. A special thank you to my wife Dr. Amanda Fixsen for her unwavering encouragement and help maintaining a work-life balance.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Table Page

CHAPTER I

INTRODUCTION

School-Wide Positive Behavior Interventions and Supports (SWPBIS) is a systems intervention that has been used by over 19,000 schools in the United States focused on creating safe and healthy school environments by encouraging pro-social behavior and providing support for students with behavioral needs (Horner, July, 2013; *School-wide PBIS: What is school-wide PBIS?*, 2013). SWPBIS has been documented to decrease office discipline referral and suspension rates, increase on-task behaviors, and improve academic performances (Algozzine & Algozzine, 2007; Bradshaw, Mitchell, & Leaf, 2010; Horner et al., 2009; Luiselli, Putnam, Handler, & Feinberg, 2005). As SWPBIS is a systems intervention, implementing the program with fidelity requires teachers and administers to work cohesively. The Team Implementation Checklist (TIC) (Sugai, Horner, & Lewis-Palmer, 2002, 2009; Sugai et al., 2011) is a measure designed to monitor implementation during start-up phases by assessing program fidelity at various levels within schools. Further, the measure is self-administered quarterly throughout the school year, providing an opportunity to empirically examine fidelity growth for schools during the initial implementation process. In this study, I will use TIC scores to model fidelity growth to illustrate a fully Bayesian approach to multilevel growth modeling and will discuss associated implications. Data for this study came from five distinct samples allowing for a demonstration of both uninformative and informative prior distributions on model results. I will use a Bayesian updating (BU) process to allow for modified contextual inferences based on the inclusion of informative prior distributions for

normally distributed fixed-effect parameters. Finally, the influence of school context and sustainability covariates on model-based inferences of fidelity growth will be examined.

Bayesian Methods

Bayesian estimation methods present an alternative method for producing probability estimates based on statistical models. The algorithm for this estimation procedure has its roots in the 18th century (Bayes & Price, 1763), and modern computing has facilitated its use in many fields. In recent years, Bayesian estimation has been used in everything from measurement research (e.g., Fukuhara & Kamata, 2011) to statistical analysis of single case design (e.g., Shadish, Rindskopf, Hedges, & Sullivan, 2013) to updating search algorithms for planes lost at sea (e.g., Caudle, 2010). Part of the appeal of Bayesian estimation is that it facilitates estimation for complex models, and another part of the appeal is in the way results are presented. As opposed to traditional least squares and likelihood estimators where results are reported in the form of a point estimates with a standard error, Bayesian estimates are reported in the form of a probability distribution.

The main concern with Bayesian estimation is philosophical. Beginning in the early 20th century, the logic of Bayes theory was questioned and even overtly criticized by notable statisticians including Fisher (Andrews & Baguley, 2013). What is troubling is that the Bayesian estimation algorithm incorporates both the data distribution as least squares and likelihood estimators do, and a prior distribution. The prior distribution can be completely vague and have relatively little influence on results, or it can be very specific and weight results. While some have argued that the incorporation of prior knowledge into the estimation algorithm makes it an inductive process, recent

interpretations argue that it can be a deductive process through the use of model checking (Gelman & Shalizi, 2013). This opinion is succinctly stated: prior distributions "are just assumptions of our model. Like any other assumptions, they can be good or bad and may need to be extended, revised, or possibly abandoned on the basis of their suitability to the data being studied" (Andrews & Baguley, 2013, p. 6).

Multilevel Linear Growth Models

Multilevel/hierarchical linear models provide one method for analyzing longitudinal change in an outcome. By isolating variance associated within and between individual units, these models are used to create statistical estimates of growth (Raudenbush & Bryk, 2002). Multilevel linear models have been used to in many instances to understand student level change and school variables associated with that change (e.g., Biancarosa, Bryk, & Dexter, 2010; Bryk & Raudenbush, 1988; Zvoch & Stevens, 2006). Multilevel linear growth models have also been used to understand organizational change (e.g., Bradshaw, Koth, Bevans, Ialongo, & Leaf, 2008) and increased programmatic fidelity (e.g., Horner et al., 2009). Three distinct estimation options for multilevel linear models are described.

Maximum likelihood estimates. Maximum likelihood (ML) estimators are probably the most used estimator for multilevel linear models. ML estimates are based on an joint data distribution of the observed data included in the analysis and often assume multivariate normality (Enders, 2010; Gill, 2002; Raudenbush & Bryk, 2002). Based on an iterative process, ML procedures maximize the likelihood for the observed data given parameter estimates.

Fully Bayesian estimates. Bayesian estimates account for the likelihood distribution and a prior distribution. Because results are reported as probability distributions, some prefer Bayesian approaches for multilevel modeling (e.g., Gelman, Hill, & Yajima, 2012). The prior distribution can be uninformative and resulting estimates are based solely on the data. When uninformative prior distributions are applied, the results are approximately the same at the ML estimates (Andrews & Baguley, 2013; Gill, 2002). Prior distributions can be informative, and subsequently influence the results based on their shape and size. The use of prior distributions can be criticized as challenges easily arise as to what constitutes prior information (Gill, 2002), and how prior knowledge is transformed into explicit probabilities with distributional form and size (O'Flaherty & Komaki, 1992).

Bayesian updating. Bayesian updating (BU) provides a method for explicitly incorporating informative prior distributions based on previous sample results. In a sense, this takes some of the guesswork out of specifying prior distributions as results from previous analyses using Bayesian estimators are already in a probabilistic form of a posterior distribution. In turn, posterior distributions can then be applied in serial fashion as prior distributions into the analyses of subsequent samples. BU methods have been applied in many fields ranging from measurement research to synthesizing findings collected over several studies (Kuiper, Buskens, Raub, & Hoijtink, 2013; Zwick, Ye, & Isham, 2012). BU methods have also been shown to improve overall model fit (Yu & Abdel-Aty, 2013).

Model Selection

In addition to using informative priors from previous samples, model checking can facilitate an objective Bayesian estimation process (Gelman & Shalizi, 2013; Morey, Romeijn, & Rouder, 2013). A method to facilitate this process is to conduct a sensitivity analysis where models with varying specification differences are compared to one another using model selection criteria (Gill, 2002). Bayes factors and information criteria provide numerical indices facilitating objective comparison between models to compare local (i.e., how well the model fits the data) and global (i.e., how well the model predicts future data) model generalizability (Liu & Aitkin, 2008).

Study Aim

The aim of this study was to demonstrate a fully Bayesian approach to model multilevel linear growth. Further, the extent that informative distributions produced through a BU process influence parameter estimates was explored. Model selection criteria including approximate Bayes factors, AIC, BIC, and DIC were used to compare Bayesian models with uninformative and informative prior distributions. Models used in this demonstration estimate within-year fidelity growth of SWPBIS. To this end, the following research questions were addressed:

1. What were the similarities and differences of within-year SWPBIS fidelity growth estimates for the 2008-9 sample using a maximum likelihood (ML) estimator and Bayesian estimator with uninformative prior distributions?
2. What are the similarities and differences of within-year SWPBIS fidelity growth estimates for the 2009-10, 2010-11, 2011-12, and 2012-13 samples when utilizing

ML estimators, and Bayesian estimators with uninformative and informative priors?

3. What is the impact of the number of years implementing, school context, and sustainability covariates on the various estimates of within-year SWPBIS fidelity growth?

CHAPTER II

LITERATURE REVIEW

The Bayesian Analytic Approach

Bayesian estimation offers a unique opportunity for analysts to incorporate prior information and facilitate estimation of complex models that may otherwise not be possible. Applied uses of Bayesian methods have shown that these techniques are another tool for analysts to use when the need arises, and the rigid theoretical boundaries between frequentists and Bayesians are being blurred. In fact, Howard Wainer (2010) strongly recommends the study of Bayesian methods stating

> Facility with them is a must for anyone who intends to make contributions to measurement in the future. And so, if the concepts associated with such terms as conjugate prior, jumping kernel, inverse-gamma distribution, and the Metropolis-Hasting algorithm are not close to your soul, get busy (p. 7).

Bayesian versus frequentists paradigms are sometimes characterized as inductive versus deductive approaches (Efron, 2005; Gill, 2002). Bayesians conceptualize probability as something that is based both on observation and existing knowledge. In fact, Bayes' theorem (Bayes & Price, 1763) incorporates existing knowledge and the joint distribution of observed data into a mathematical algorithm, a topic that is discussed further. Frequentists, alternatively, conceptualize probabilities deductively and defend their approach as primarily objective as estimates are based only on observed data. Luckily for the modern analyst confronted by larger and larger amounts of data complexity, the predominant opinion is nicely summarized: "Bayes rule is a very

attractive way of reasoning, and fun to use, but using Bayes rule doesn't make one a Bayesian. *Always* using Bayes rule does" (Efron, 2005, p. 2).

The theoretical divide between frequentist and Bayesian paradigms might not even be very large. Gelman and Shalizi (2013) argued frequentist approaches are not fully objective and Bayesian approaches are not inductive. The authors claim while frequentists estimates are based solely on data, statistical modeling is based on many subjective decisions by researchers and analysts thus negating its objectivity. Model selection, for example, can be defended with hypothesis testing and various information criteria, but the choice of what models to compare is a choice made by the analyst. Interestingly, the authors point out that Bayesian approaches are not inductive, but Bayesian modeling is better characterized as *hypothetic-deductive* emphasizing model analyses do not end with estimating parameter posterior distributions. The authors claim that Bayesian inference is better characterized as *deductive* because of using observed distributions, and modeling decisions that can be falsified using model tests. It is *hypothetic* because hypotheses can be explicitly incorporated into an analysis with the inclusion of prior information. In the end, an analyst must consider multiple statistical models, assume all statistical models are false, and make decisions, as objectively as possible, between competing models that—imprecisely—model the phenomenon of interest.

In the modern scientific era, statistical analysts are confronted with growing data complexities such as multiple levels, randomization of experiments at higher level clusters, second and third order factorials, and the list goes on and on. When absorbed in the minutia, it is often difficult to maintain focus on project goals. Hughes (1997) offers

a simple conceptualization of scientific modeling: Denotation, demonstration, and interpretation. Denotation is the process where pieces of the real world are denoted by elements within the model; the model dynamically demonstrates theoretical conclusions; and finally is interpreted to explain and predict phenomenon of interest. Use of a Bayesian approach to statistical modeling allows for the explicit incorporation of prior information, ranging from completely unknown to very specific, into denotation, demonstration, and interpretation of the model through the use of Bayes Theorem and it's adapted estimation algorithm.

Bayes theorem and Bayesian estimation. Richard Price posthumously published Thomas Bayes' probability theorem in the eighteenth century (Bayes & Price, 1763). The theorem allows for the estimation of a posterior parameter distribution given a prior distribution and the observed data. Specifically, the theorem is often characterized as an inverse probability and is expressed (Larsen & Marx, 2001)

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_{j=1}^{j} P(B \mid A_j)P(A_j)}, \tag{1}$$

where the probability of A in the state *j* given B equals the probability of B given A_j times the probability of A_j divided by the sum of the probability of B given all, *J*, states of A times the probability of A. This expression is known as the inverse probability because it allows for the calculation of the probability of a state, A, given the contingency of another state, B. Further, the calculation is based on knowing the inverse, the probability of state B given state A, and the probability of A independently.

The following example paraphrased from a statistical textbook highlights the inverse quality of the theorem (Larsen & Marx, 2001, p. 50). Suppose that 100 people

are arrested in suspicion of looting during a blackout and given a polygraph. It is known that there are only 12 guilty suspects out of the 100 arrested, giving the probabilities $P(A_{guilty}) = .12$ and $P(A_{innocent}) = .88$. Further, it is assumed that polygraphs are 90% reliable when administered to a guilty suspect and 98% reliable when administered to an innocent subject (or put conversely, 2% of the time polygraph results will imply guilt when a person is innocent). Let's say B represents a guilty polygraph result and A_j represents whether a person is guilty or innocent. Using this parameterization, the $P(B_{guilty}|A_{guilty}) = .90$ and $P(B_{guilty}|A_{innocent}) = .02$. Using this information and Equation 1 allows for the computation of the probability that a person is guilty given a positive result from the polygraph test.

Given that we know the number of states, $J = 2$, of A and all the other relevant information we can substitute values into the formula as follows:

$$P(A_{guilty} \mid B) = \frac{.90*.12}{(.90*.12) + (.02*.88)} = \frac{.1080}{.1080 + .0176} = \frac{.1080}{.1256} \approx .8599. \quad (2)$$

These results imply that given a guilty polygraph result, a suspect has about an 86% change of being guilty, or conversely a 14% chance of being innocent. This exercise highlights how the probability of a suspect guilt or innocence is contingent upon the inverse probability. That is, knowing what the probability of a guilty positive polygraph result given a suspect's guilt or innocence allowed the computation of the desired probability.

Being able to calculate the inverse probability given the conditional probabilities has interesting ramifications for probability exercises, but the real utility is the expansion of this probability theory into an estimation algorithm. While the following theorem is in

some ways much more complex than other estimation algorithms, it allows for simple interpretation of parameter results. The Bayesian estimation algorithm is as follows (Gelman, Carlin, Stern, & Rubin, 2004; Gill, 2002; Larsen & Marx, 2001):

$$f(\theta|X) \propto L(X|\theta)g(\theta), \tag{3}$$

where $f(\theta|X)$ is the posterior density function of the parameters, θ, given the data, X; $L(X|\theta)$ is the likelihood function; and $g(\theta)$ is the prior parameter distributions. The complexity in using Bayesian methods arises as (a) the solution to the estimation algorithm is a density function where parameter estimates are distributions themselves rather than point estimates with an error term, (b) estimation requires specification of a prior distribution for the parameters, and (c) using the approach requires that the analyst understand the mathematical relationship between the likelihood function and the prior distribution. The simple interpretation is a result of parameters being expressed as a distribution with a credibility interval rather than a point estimate with a standard error that is based on sampling theory. Put another way, estimates lie within a calculated range (i.e., a posterior distribution) rather than a specific number with theoretical error that would occur over repeated samples.

For example, suppose we are estimating the mean difference between an experimental and control group for a randomized trial. Frequentist approaches would allow us to calculate the mean difference and its standard error that could then be used to make an inference based on the *t* or *Z* distribution. The standard error implies the variation in the mean difference that would theoretically occur over repeated sampling from the same population, and inference on the mean difference is contingent upon the analyst's choice of a *t*- or *z*-test. That is, the inference of whether the treatment had an

effect is contingent upon the choice of the statistical test and theorized variation over many samples. Results in the frequentist tradition are reported as a point estimate of the mean difference and error surrounding that estimate. If the same example was estimated with a Bayesian approach the result, the mean difference between the treatment and control groups, would be a distribution itself and results would be reported within a range rather than a point estimate with error. The simplicity of the Bayesian results is that the statistical estimate, the posterior distribution, is a range that explicitly incorporates the uncertainty.

Confusion can easily arrive, however, as Bayesian estimates do not utilize *p*-values because estimates are in the form of a distribution. In the previous example, the statistical test based on the point estimate and standard error allows an analyst to assign probability to the test and compare results to a threshold. For *z*- and *t*-tests with larger samples, this threshold is about +/-1.96 for a two-tailed hypothesis test at a threshold of .05 for *p*. If the statistic is beyond the limit of this threshold, the test can be said to be statistically significant. Uninformed consumers of research often are primarily concerned with this result, and the absence of such conventions when using Bayesian approaches may be unappealing. Kruschke (2013), however, points out the interpreting *p*-values is a complicated process based on sampling intention and distributional properties of observations, and "basing conclusions on "the" *p* value and "the" significance is a misleading ritual" (p. 590). To the informed, the reporting of the range of possibility for parameter estimates succinctly and elegantly demonstrates the probabilistic nature of statistics.

While the elegance of interpreting results is appealing, further discussion of the complexities of using a Bayesian approach must be discussed. A Bayesian approach to statistical modeling follows similar patterns as more traditional statistical modeling including specifying an analytic model, declaring and evaluating assumptions, sensitivity analyses, and comparing alternative models (Gelman & Hill, 2007; Gill, 2002; Kuiper et al., 2013). Special considerations for Bayesian modeling includes specifying prior distributions, specifying the form of posterior distributions, posterior predictive checking, comparing alternative models using a Bayes factor, and Gibbs sampling. While the full details of these techniques will be outlined in the methods section, the underlying requirement is that the analyst must possess knowledge of the parameters of interest and mathematical understanding of the estimated joint density function of the statistical models being utilized.

Like all analyses, models must be continuously checked in terms of mathematical and specification assumptions. This is especially important when using Bayesian estimators, as additional assumptions such as the shape and size of the prior distribution are explicitly included in the model. Subsequently, continuous model checking is a must (Gelman & Shalizi, 2013; Morey et al., 2013).

> The data-analysis process – Bayesian or otherwise – does not end with calculating parameter estimates or posterior distributions. Rather, the model can then be *checked*, by comparing the implications of the fitted model to the empirical evidence. One asks questions such as whether simulations from the fitted model resemble the original data, whether the fitted model is consistent with other data not used in the fitting of the model, and whether variables that the

> model says are noise ('error terms') in fact display readily-detectable patterns. Discrepancies between the model and data can be used to learn about the ways in which the model is inadequate for the scientific purposes at hand, and thus to motivate expansions and changes to the model (Gelman & Shalizi, 2013, p. 12).

Model selection. Model selection is an important analytic decision especially given that the choice of one model over other could alter inferences based on model parameters. A helpful way to approach model selection is the concept of generalizability, which put succinctly is a model's ability to predict future data (Liu & Aitkin, 2008). To this end several options exist for model selection including deviance tests, information criteria, and Bayes factors.

Model change deviance tests involve using a test statistic, usually a chi-square, to determine if the change in deviance from one model to another is significant. Deviance is typically defined as negative two times the log-likelihood, and the degrees of freedom for a change in deviance test is the difference in number of parameters between one model and another (Gelman & Hill, 2007; Raudenbush & Bryk, 2002). For Bayesian models, many computer programs including the JAGS program (Plummer, 2003) report deviance in terms of a posterior distribution where the mean deviance is equal to the deviance plus the number of predictors. In all cases the lower the deviance implies a better fitting model, and subsequently higher level of generalizability.

Information criteria are based on deviance information and additional model information in the calculation of an adjusted value. The Akaike information criterion (AIC) for example is the deviance plus two times the number of predictors (Gelman & Hill, 2007). The Bayesian information criterion (BIC) includes an adjustment for sample

size and is the deviance plus the number of parameters times the natural log of the sample size (Liu & Aitkin, 2008). Finally, for Bayesian models the deviance information criterion (DIC) is considered the Bayesian equivalent to the AIC and is calculated by adding the mean deviance and effective number of parameters, where the effective number of parameters takes into account pooling for multilevel models (Gelman & Hill, 2007; Liu & Aitkin, 2008; Spiegelhalter, Best, Carlin, & van der Linde, 2002). Unfortunately, calculating the effective number of parameters is unstable so caution should be used when interpreting results (Gelman & Hill, 2007, p. 525). Because AIC, BIC, and DIC are based on the deviance, a lower value is considered a better fitting model.

Specific for Bayesian models is the use of a Bayes factor. A traditional Bayes factor is the ratio of posterior probabilities for competing models and involve the computations based on probability density functions and prior distributions (Gill, 2002; Liu & Aitkin, 2008). Basically, the Bayes factor provides evidence of one model over another given the data at hand. An alternative to the traditional Bayes factor is the approximate Bayes factor based on the BIC and is calculated using Equation 4 (Liu & Aitkin, 2008, p. 365):

$$B_{21} = \exp[-\frac{1}{2}(\mathrm{BIC}_2 - \mathrm{BIC}_1)]. \quad (4)$$

Given the simplicity of calculation, the approximate Bayes factor might be preferred in many situations. The result of calculating Bayes factors is a ratio that if greater than one supports the second model over the first, and if less than one does not support the second model (Jeffreys (1961) as cited in Gill, 2002, p. 242).

Given the multitude of methods for choosing models, I agree with Liu and Aitkin's (2008, p. 365) recommendation that multiple criteria should be presented for model comparison. If the criteria converge on one model in favor of the other more evidence for that model has been presented. For example, if one model has the lowest values for AIC, BIC, and DIC and the approximate bayes factor for that model compared to a competing model is greater than one, then the researcher has ample evidence for choosing one model over the competing model.

Applied use of Bayesian methods. Bayesian methods are becoming more and more abundant in various areas of educational and psychological research. Measurement topics such as Item Response Theory (IRT) provide numerous examples as models require estimation of many parameters. Interestingly, the analysis of single case design has also taken a vested interest in Bayesian methods as they pose potential solutions to modeling small sample data. Finally, as more people have become familiar with complexities in statistical tests, new approaches to hypothesis testing are being explored.

Measurement research. IRT applications of Bayesian methods are perhaps the earliest examples to appear in the educational and psychological literature. Beginning in the 1980s, simulation studies demonstrated the strength of Bayesian estimation algorithms over maximum likelihood in various scenarios when estimating Rasch models (Swaminathan & Gifford, 1982), two parameter logistic (2PL) models (Swaminathan & Gifford, 1985), and three parameter logistic (3PL) models (Swaminathan & Gifford, 1986). These results were replicated with moderately small samples (Mislevy, 1986). In the 1990s IRT researchers used Bayesian methods to improve estimates of score locations for binary and partial credit items (Huynh, 1998) and estimation with a marginal

Bayesian approach (Zeng, 1997). More recently, Bayesian methods have been used in increasingly complex models like the logistic positive exponent model where item characteristic curves are asymmetrical (Bolfarine & Bazan, 2010). Also, methods for analyzing dependencies between items and persons found in testlet data have been developed using Markov Chain Monte Carlo (MCMC) methods (Jiao, Kamata, Wang, & Jin, 2012), and applied to improving Differential Item Functioning (DIF) detection for testlet items (Fukuhara & Kamata, 2011).

Noting dependencies and cross-loadings in measurement models, Muthén and Asparouhov (2012) proposed an intermediary process between exploratory and confirmatory factor analysis. By specifying prior distributions for covarying and cross-loaded items using a conjugate prior (in this case, the inverse Wishart), the authors showed how unresolved dependencies can be explicitly modeled. In doing so, models that may be non-identified in a traditional structural equation model framework can be estimated[1].

In another avenue of measurement research, DIF methods using Bayesian approaches have been explored. For example researchers have developed the following models using Bayesian approaches: A bifactor multidimensional IRT model for DIF detection on testlet items (Fukuhara & Kamata, 2011), a model for simultaneous detection and explanation of DIF (Soares, Gonçalves, & Gamerman, 2009), estimation of DIF in small samples sizes (Sinharay, Dorans, Grant, & Blew, 2009), and applying Bayesian updating (BU) to Mantel-Haenzel estimates (Zwick et al., 2012). All and all, this work provides evidence in the validity of the technique in complex statistical models.

[1] Bayes nets offer another alternative for modeling dependencies not offered with other techniques (Almond, Mulder, Hemat, & Yan, 2009).

Statistical analysis of single case design. Statistical methods for single case design data have focused on how to model small sample data to provide numerical support for visual analysis. Considering the small sample size for single case research and the relatively few number of data points to model trend using autocorrelation methods, Shadish, Rindskopf, Hedges, and Sullivan (2013) demonstrated that Bayesian methods compensate for sampling error associated with other estimators. To assess intervention impact, de Vries and Morey (2013) introduced methods for conducting hypothesis tests using the Bayes Factor that can account for dependencies resulting from longitudinal data collection. As with measurement research, these examples provide evidence for applying Bayesian methods when other methods do not suffice.

Hypothesis testing. A central facet of social science research is hypothesis testing. Specifically, hypothesis testing allows researchers to specify a null hypothesis, and reject or retain that statement based on a statistical test. A central problem with traditional (frequentist) hypothesis testing is "the fact that the resulting probability value does not tell the researcher what he or she usually wants to know: How probable is a hypothesis, given the obtained data?" (Masson, 2011, p. 679). Further, hypothesis testing is used for everything from testing single parameter models (i.e., the *t*- or *z*-test) to significance tests within larger models (i.e., using a *t*-test to determine if a parameter is significant within a regression model) to testing differences between plausible models (i.e., conducting a χ^2 test on the change in deviance between nested models). Coupled with these diverse uses of hypothesis testing are complications arising from model assumptions. Because of these challenges, Kruschke (2013) claimed that Bayesian methods for hypothesis testing may be more appropriate than traditional *t*-tests. In fact, using a Bayes Factor allows for

testing multiple hypotheses simultaneously (Gill, 2002) that can be used for as an analog for traditional hypothesis testing, constrained hypothesis testing (Klugkist, Laudy, & Hoijtink, 2010; Morey & Rouder, 2011), and comparing models (Gelman et al., 2004).

Other examples. The rise in popularity of Bayesian methods is evident in many other ways. The number of computer tools developed for specific Bayesian approaches has risen dramatically (e.g., Campbell & Thompson, 2012; Morey & Morey, 2011; Vanpaemel, 2009; Wetzels, Lee, & Wagenmakers, 2010). Besides the aforementioned methods, developments have been made using Bayesian approaches for many other specific applications including estimating polychoric correlations (Choi, Kim, Chen, & Dannels, 2011), sample size estimations for cluster randomized trials (Rotondi & Donner, 2009), growth mixture modeling (Depaoli, 2013), mediation analysis (Yuan & MacKinnon, 2009), and even rater evaluations for grant reviews (Cao, Stokes, & Zhang, 2010). In sum, researchers and analysts are realizing the potential of applying Bayesian methods and testing the methodological feasibility of applying these methods within diverse contexts.

Bayesian Updating. While the use of Bayesian methods may be complex, they offer opportunities not afforded by other techniques such as the process called Bayesian updating (BU). BU allows for the explicit incorporation of a previous trial results into the analysis of subsequent results. Specifically, the posterior distribution of a previous trial is incorporated as the prior distribution for the subsequent trial. Gill (2002) notes "this cycle of prior to posterior is actually a very solid way of conceptualizing the scientific process: we take what knowledge we have in hand and update it with new information when such results become available" (p. 72).

Figure 1 depicts an idealized view of Bayesian inference using the updating process. The image borrowed from Gelman and Shalizi (2013, p. 9) shows that as time moves forward, the probability of the previous model decreases as the probability of the subsequent model increases. This occurs as new information is gathered about the phenomenon of interest and models can be updated using that information. From an idealized perspective, this is an inductive process as denotations, demonstrations, and interpretations (Hughes, 1997) of models of reality are revised as new information is gathered. As previously discussed, Gelman and Shalizi (2013) highlight that this is not an inductive process, but a hypothetic-deductive one guided by many decisions of the analyst.

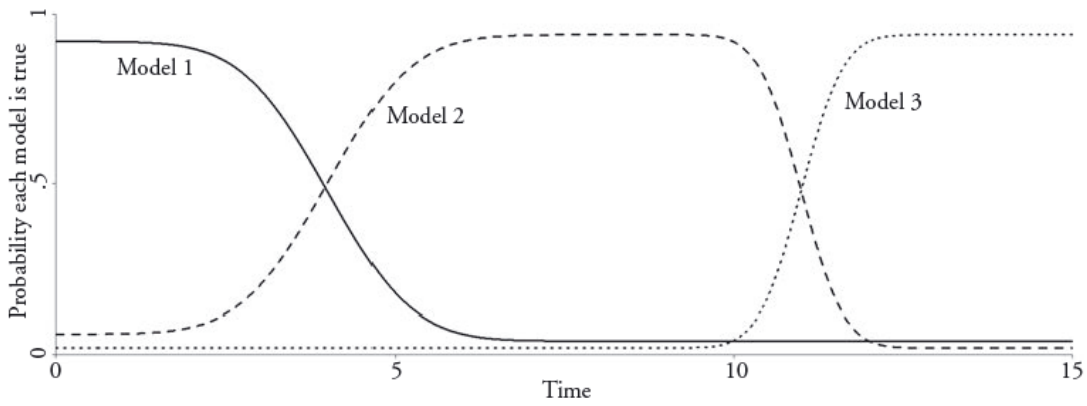

Figure 1. An idealized view of Bayesian inference using Bayesian updating (reproduced from Gelman & Shalizi, 2013, p. 9). The posterior probability of each model changes over time as estimates are updating using posterior results from the previous model as the prior distribution for the subsequent model.

The analyst's decision-making process is further complicated by many considerations. Inferences based on estimates produced using updated priors are

contingent on data sets being independent and identically distributed (a mathematical property often written as *iid*) and from the same sampling plan or data generating process as the data sets used to estimate prior distributions (Gill, 2002, p. 72). If the structure of the statistical model does not change from one data collection wave to another, the results from the second wave will be nearly identical to the likelihood estimates of the pooled data across waves. The structure of the model may change, however, as new information is learned about the phenomenon being modeled and incorporated into the specifications of prior distributions of subsequent analyses.

It is this characteristic that makes Bayesian methods a unique and valuable tool for analysts. Being able to adapt statistical models and incorporate prior information embeds analyses within a scientific process that occurs over time. Further, competing models[2] can be compared objectively using techniques such as the Bayes factor (Klugkist et al., 2010; Morey & Rouder, 2011), a topic further discussed in the methods section.

Applied use of Bayesian Updating. Philosophical debates aside, the practical process of BU is characterized by O'Flaherty and Komaki (1992) as the specification of 'warranted probabilities' based on 'warranted beliefs'. That is, previous research and theory provide the basis for beliefs that then need to be transformed into probabilistic statements. In order to do this, analyst must also use specialized software such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), JAGS (Plummer, 2003), the Bayes estimator in Mplus (L. K. Muthén & Muthén, 1998-2012), and/or a new program called Stan (The Stan Development Team, 2013).

[2] This highlights one reason why Gelman and Shalizi (2013) claim that Bayesian approaches are deductive because models can be compared using a statistical test.

Building on early applications of Bayesian methods in the Item Response Theory (IRT) field (i.e., Mislevy, 1986; Swaminathan & Gifford, 1982, 1985, 1986) several researchers applied BU to various models. BU has been used to estimate item information for three-parameter logistic (3PL) and monotone partial credit models (Huynh, 1998). 3PL models are often difficult to estimate using maximum likelihood techniques (Embretson & Reise, 2000) as the likelihood surface is very flat for the third parameter and searching for a maximum often results in non-convergence. Subsequently, BU has been used with marginal estimation of three-parameter models by incorporating empirical means as prior distributions (Zeng, 1997).

More recently within the field of measurement, BU has been applied to differential item functioning (Zwick et al., 2012). Within a simulation study, Mantel-Haenzel estimates were more stable when incorporating information from previous test administrations as prior distributions. Further, the BU approaches appeared more stable than previous attempts using empirical Bayes methods.

Other areas are beginning to apply BU methods. These methods have been used for studying economic risk based on natural disasters (Botzen & van den Bergh, 2012; Kelly, Letson, Nelson, Nolan, & Solís, 2012), road accident prevention (Deublein, Schubert, Adey, Köhler, & Faber, 2013), and even for updating search algorithms for ships and planes lost at sea (Caudle, 2010). Kuiper, Buskens, Raub, and Hoijtink (2013) demonstrated in a simulation study how evidence from several studies that might otherwise confound meta-analytic approaches because of different dependent and predictor variables used to measure same theoretical concepts, could be accurately estimated using updated priors. Coupled with this work, is research focusing on the

impact of the form of the informative prior distribution on posterior distributions estimates demonstrating that use of BU improved overall model fit and parameter accuracy (Yu & Abdel-Aty, 2013).

The growing empirical base demonstrating the use of Bayesian methods, BU, and methodological work refining understanding of these techniques provides guidance for using these techniques in complicated scenarios. The aim of this study is to demonstrate how analyst can apply BU methods to serially update prior distributions in a multilevel linear growth model. Specifically, fidelity growth of a school universal behavior intervention will provide the context for demonstrating the methodological technique.

Treatment Fidelity

In the last decade, O'Donnell (2008) showed that few educational research articles documenting intervention impacts measured intervention fidelity and analyzed fidelity's association with outcomes. Even in high impact special education journals, only 67.4% of studies using group design reported collecting fidelity data between 2005 and 2009 (Swanson, Wanzek, Haring, Ciullo, & McCulley, 2013). To further complicate the matter, Hulleman and Cordray (2009) documented a decrease in treatment fidelity when an intervention was implemented in a classroom setting as compared to a laboratory setting highlighting the need to understand fidelity in school-based trials. In response to these challenges, educational researchers have looked to other fields for guidance on how to create models of treatment fidelity, test strategies for enhancing fidelity, and create and validate measures of fidelity (e.g., Schulte, Easton, & Parker, 2009). Subsequently, many education researchers are beginning to measure fidelity in a multi-faceted fashion.

For example, researchers focusing on the implementation of a classroom management system reported collecting data at many levels and occasions including workshop training sessions, regular coaching meetings, and during classroom delivery (Reinke, Herman, Stormont, Newcomer, & David, 2013). This exploratory study found (a) trainers delivered workshops with fidelity, (b) coaches implemented the coaching model with fidelity, (c) workshops leaders' engagement ratings were associated with teacher implementation of the praise-based classroom management system, and (d) different levels of coaching were associated with teacher implementation and student behavior outcomes. Even from an exploratory perspective, the importance of documenting the interconnected nature of fidelity manifests itself.

In a different fashion, researchers exploring the relationship between fidelity of implementation and student outcomes of a computer-based middle school math curriculum measured fidelity using two overarching fidelity constructs (Crawford, Carpenter, Wilson, Schmeister, & McDonald, 2012). The authors defined (a) *fidelity to structure* as time spent on the intervention, teacher adherence to the program, and student engagement; and (b) *fidelity to process* in terms of variables essential to computer based instruction including teaching communication, classroom management, and problem-solving skills. In this single group pre-post test study of 654 seventh and eighth grade students and 23 teachers in 11 public schools spread across seven states, hierarchical linear models controlling for pre-test revealed that *fidelity to structure* variables had a significant impact on student outcomes while *fidelity to process* variables did not.

Focusing on reading intervention for middle and high school students, a team of researchers explored the relationship of implementation fidelity and teacher efficacy on

student reading achievement (Cantrell, Almasi, Carter, & Rintamaa, 2013). For this study, the authors defined implementation fidelity as teachers' adherence to critical components of the intervention curriculum, and efficacy as a teacher's belief that she/he can influence desired outcomes. In this exploratory study of nine sixth grade and 11 ninth grade teachers, results revealed that teacher efficacy was related to student's reading comprehension and overall achievement, while fidelity was related to student's growth in vocabulary.

School-level interventions' influence on distal outcomes. Linking interventions delivered to school and district personnel to student outcomes provides documentation between systems level interventions and valued results. Studies have shown that interventions delivered at the school and teacher levels can impact student outcomes. For example, a model of one-on-one coaching of literacy skills enhanced student outcomes during implementation (Biancarosa et al., 2010) documenting the relationship with treatment delivered at the teacher level and outcomes at the student level. Similar initiatives delivered to school and district personnel focused on enhancing data-use to meet students' behavioral and academic needs were associated with enhanced student performances (Carlson, Borman, & Robinson, 2011; Chaparro, Smolkowski, Baker, Hanson, & Ryan-Jackson, 2012).

Analyzing how varied implementation levels impact student outcomes using statistical models, however, is a difficult and often unfruitful task. Including an implementation or fidelity variable as a school-level covariate in multi-level models evaluating school level interventions such as literacy initiatives on student growth can lead to non-significant results (Zvoch, Letourneau, & Parker, 2007). This may be due to

the complexity of implementation as evidence has shown that teacher, classroom, and school-site characteristics were associated with program fidelity (Zvoch, 2009), and treatment fidelity may better be conceptualized as a "multilevel, multidimensional construct" (Zvoch, 2012, p. 558). To better understand the implementation process, a framework for implementation is outlined.

A Framework for Implementing Evidence-Based Programs

Documenting the impact of a practice on valued outcomes is one piece of essential evidence for scaling educational innovations. Another critical piece is an understanding how to effectively put a practice into place. *Implementation* and *scaling-up* are two terms often associated with installing research-based innovations within real-world organizations. This is no easy task as Coburn (2003) highlighted that to understand how to make "deep and consequential change" (p. 4) requires assuming "the problem of scale is fundamentally multidimensional" (p. 3), rather than just the sheer number of organizations using an innovation. To this end, Fixsen, Naoom, Blase, Friedman, and Wallace's (2005) framework of implementation based on a large literature base will be utilized to frame the functional steps of putting a program into practice and organizational supports necessary for promoting consistent use of new practices. The functional steps for installing an evidence-based program is outlined in the *stages of implementation* section, and the organizational supports needed are described in the section on *implementation drivers*.

Stages of implementation. Fixsen and colleagues (2005) described the stages of implementation as a guide for how to install innovations within organizations based on a systematic review of literature focused on implementation. The stages are *exploration*,

installation, *initial implementation*, and *full implementation*. During the *exploration* stage, an organization's readiness is assessed to determine if an organization has the capacity to implement a program and its fit within the context of the organization. The *installation* phase is characterized by preparing necessary functions within the organization for the new program. A range of actions occurs during this stage from selecting staff to acquiring necessary office space. *Initial implementation* is when a new practice is used for the first time. This can be quite a challenge as practitioners use newly acquired skills and tools for the first time, and success relies heavily upon implementation drivers such as training, coaching, and ongoing administrative support. Finally, *full implementation* is defined as having more than 50% of program staff using the innovation with fidelity and producing improved outcomes.

Implementation drivers. Fixsen and colleagues (2005) described the integrated drivers that facilitate consistent and effective use of an innovation within an organizational setting. This model is depicted in Figure 2. The drivers were divided into three main categories: *competency*, *organization*, and *leadership*. *Competency* drivers refer to coaching and training. *Organizational* drivers refer to systems interventions, facilitative administration, and decision support data systems. Finally, *leadership* drivers refer to the technical and adaptive characters of management. Further, the individual drivers and three main domains are integrated such that they correlate, interact, and depend upon one another. For example, data systems can enhance organizational drivers and are reliant upon personnel competency to access, interpret, and use data for decision-making purposes.

.

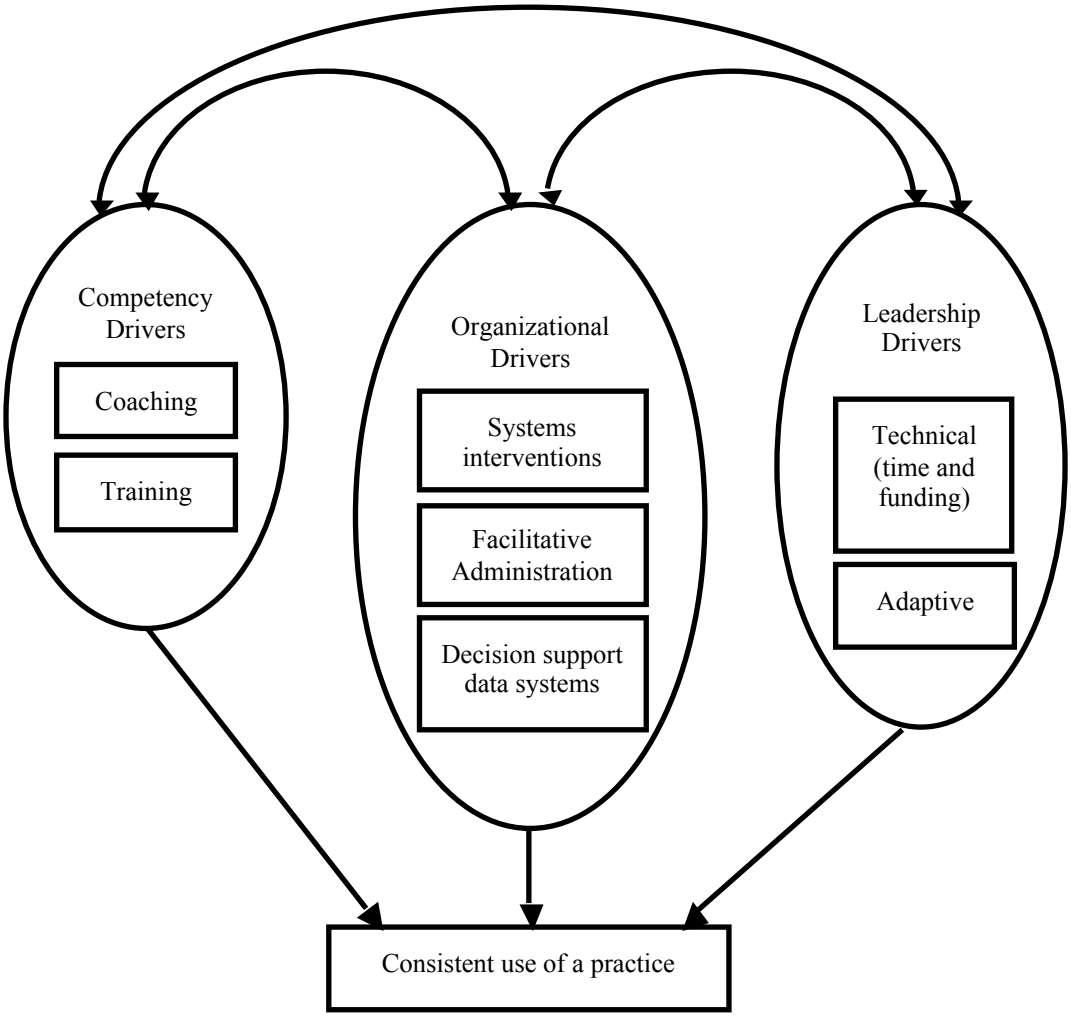

Figure 2. An integrated view of organizational drivers that facilitate consistent and effective use of innovations (adapted from Fixsen et al., 2005).

Research on implementation drivers. One way to approach implementation research is through a systemic line of work focused on building an evidence base for the individual drivers involved. To this end, data based decision making research has documented the relationship between implementation drivers (e.g., decision support data systems, adaptive leadership, and training) and valued outcomes. For example,

consistent use of data has been associated with improved behavior outcomes in schools using school-wide positive behavior interventions and supports (SWPBIS) (Ervin, Schaughency, Matthews, Goodman, & McGlinchey, 2007). Further, effective data use was related to the data's accessibility, timeliness, and perception of validity, as well as training and support for data analysis (Kerr, Marsh, Ikemoto, Darilek, & Barney, 2006) highlighting the importance of competency and organizational drivers. Focused solely on training, effectiveness trials have documented training delivered by research teams and increased use of data based decision making protocols (Newton, Horner, Algozzine, Todd, & Algozzine, 2012; Todd et al., 2011). In turn, Newton and colleagues (2011) showed how this training could be brought to scale by demonstrating that their research team could teach trainers how to deliver the training, and these newly minted trainers could deliver the training to school teams with fidelity. As a whole, these studies provide evidence for interventions enhancing important implementation drivers.

Focusing on individual implementation drivers, however, might be misleading to the uniformed consumer of research. Documenting the relationships of individual implementation drivers provides an evidence base for an implementation framework, but perhaps overlooks the multidimensionality of effectively installing new practices. Given that implementation is a longitudinal process, measurement and evaluation focusing on implementation fidelity growth is pertinent.

Measuring implementation fidelity growth. Implementation fidelity growth can be used to link the longitudinal process for implementing a program and possible influential variables. For their research on class-wide peer tutoring, Buzhardt, Greenwood, Abbot and Tapia (2006) developed a *rate of implementation* scale, and

highlighted the importance of this work in relation to scaling innovations by explaining that

> A critical piece to understanding an educational intervention's scalability is knowledge of the tasks required to move from *no* implementation to *complete* implementation, how long it takes to get to that point, and the factors affecting the rate of achieving those tasks (p. 486).

To this end, the researchers developed a rate metric that measured the amount of time it took school teams to implement the critical aspects of the instructional intervention system. Employing an exploratory analysis of a wait-list control randomized trial of a classroom tutoring intervention, the authors provided a model for measuring the amount of time it took the 55 teachers in nine schools across five states to implement the 12 critical components of the intervention. Descriptive results indicated that (a) the measure accurately and objectively measured rate of completing tasks necessary for implementation of the program, (b) schools varied widely in their implementation rates, and (c) the metric allowed documentation of barriers to completing implementation tasks.

The results of Buzhardt and colleagues' (2006) research provided evidence that a fidelity growth metric can be a useful tool for researchers and practitioners. For practitioners, similar metrics on rate of implementation could be used to monitor the implementation of new programs in order to prioritize installing key components of the innovation. For researchers focused on scaling evidenced-based practices, a measure of fidelity growth can add to empiricism focused on determining factors necessary to effectively scaling programs. Perhaps a multilevel linear growth model could be used to help understand variables associated with the fidelity of implementation process.

School-Wide Positive Behavior Interventions and Supports

School-Wide Positive Behavior Interventions and Supports (SWPBIS) is a school behavior program aimed at preventing problem behaviors and creating a safe school environment (*School-wide PBIS*, 2013). SWPBIS uses preventative strategies through explicit teaching of school behaviors in various settings (e.g., classrooms, hallways, lunchrooms, etc.), and targeted interventions for students with heightened levels of behavioral needs. Given the scale of use across the country, it provides an excellent contextual example to explore fidelity of implementation growth. In the following sections, the evidence base for SWPBIS will be detailed, as well as the development of various fidelity measures.

SWPBIS impact studies. Most importantly, SWPBIS has influenced important student outcomes. SWPBIS has been documented to reduce incidences of problem behaviors as evidenced by reducing the number of office discipline referrals and suspensions (Bradshaw et al., 2010; Horner et al., 2009; Luiselli et al., 2005). Of note is a longitudinal randomized control trial of 37 schools during a 5-year period, Bradshaw and colleagues (2010) documented significant reductions in the percentage of students with office discipline referrals and overall rate of referrals, as well as reducing the rates of suspensions for treatment schools implementing SWPBIS.

Considering that low academic performance has been associated with later behavior outcomes (McIntosh, Sadler, & Brown, 2012), linking a behavior program to academic outcomes is also important. To this end, Algozzine and Algozzine (2007) documented the relationship between the use of SWPBIS and increased on-task behavior and decreased off-task behavior in classroom settings. Using a causal-comparative

research design and direct classroom observations from two schools in a large metropolitan school district, the investigators found that students answered questions with higher frequencies, spent more time talking about academics, paying attention, and raised their hands more often in the school using SWPBIS than in the schools not using SWPBIS. Further off-task behavior as defined as disrupting class, looking round, talking inappropriately, and doing an inappropriate tasks was lower in the schools using SWPBIS. In another study employing a waitlist control effectiveness trial of 30 schools receiving SWPBIS training in Illinois and Hawaii from regular state personnel rather than research staff, schools in the treatment condition were documented to improve the proportion of third graders meeting or exceeding state standards as well as increasing perceptions of school safety (Horner et al., 2009). This last study provided causal evidence for the impact of SWPBIS on academic and school-level outcomes.

SWPBIS has also been documented to influence other important school-level outcomes. For example, SWPBIS has promoted improvements in schools' overall organization health (Bradshaw et al., 2008) based on a validated measure of organizational health (i.e., the Organizational Health Inventory for Elementary Schools (Hoy & Feldman, 1987)) focused on institutional integrity, staff affiliation, academic emphasis, collegial leadership, and resource influence. Specifically, data from 2,507 staff members at 37 elementary schools participating a longitudinal group randomized study of SWPBIS revealed that while schools had similar levels prior to beginning intervention, treatment schools had significantly higher growth rates for overall organizational health, resource influence, and staff affiliation at the end of the study. Additional empirical research using cost-benefit analytic techniques has documented a positive return on

investment for time spent implementing SWPBIS on improved student behavior outcomes (Scott & Barrett, 2004). Based on a fiscal analysis of an urban elementary school with a documented SWPBIS impact of reducing office discipline referrals and suspensions over a two year period, the authors calculated that the value of the school administrators time saved was worth over $12,500 dollars or equivalent to over 30 days of administrator work over the two years.

SWPBIS measures of program fidelity. In order for evidenced-based programs like SWPBIS to work effectively, schools must implement critical program components accurately and effectively. To this end, SWPBIS researchers have created several measures focused explicitly on documenting the fidelity of critical programmatic features. The Benchmarks of Quality is a validated measure completed by school coaches and school SWPBIS team members to evaluate implementation of programmatic features in term of an item-level scale ranging from 'not in place' to 'needs improvement' to ' in place' (Cohen, Kincaid, & Childs, 2007; Kincaid, Childs, & George, 2005). A similar measure, the Team Implementation Checklist (TIC), is completed by school teams to self-assess progress towards full implementation fidelity and create implementation action plans (Sugai, Todd, & Horner, 2001). Additionally, an external evaluation tool, the School-Wide Evaluation Tool (SET), was developed and validated for yearly evaluation of SWPBIS implementation fidelity (Horner et al., 2004; Sugai, Lewis-Palmer, Todd, & Horner, 2001), and was used in the studies detailed in the following two paragraphs.

In this first example using the SET score, researchers explored the impact of SWPBIS on high school office discipline referral rates for each 100 students at

participating schools over a three year period (Flannery, Fenning, Kato, & McIntosh, in press). Using a non-randomized treatment-control design including over 36,000 students in 12 high schools, latent growth models revealed that discipline referral rates decreased for the eight schools in the treatment condition and increased for comparison schools. Further, analyses of treatment schools revealed that SET scores of program fidelity did not predict discipline referral rates during the first year of implementation, but were positively associated with decreased rates of referrals during the second and third year. This finding indicated that program fidelity had an important influence on valued-student outcomes for schools implementing SWPBIS.

In a second example, a criterion was created from the SET score to document the relationship of SWPBIS implementation fidelity and student outcomes in a state-wide program evaluation of over 400 Illinois schools (Simonsen et al., 2012). For this study, SET scores were converted to a binary fidelity criterion for meeting or not meeting a fidelity standard. Specifically, schools were labeled as meeting the fidelity standard if they met at least 80% of the fidelity criteria, and labeled as not meeting the fidelity standard if their SET score fell below the 80% level. Hierarchical linear models revealed that schools implementing SWPBIS at or above the 80% criterion had significantly lower suspension rates and higher math achievement levels on a state-mandated test.

While it is important to show that implementing with fidelity is related to outcomes, it is also important to document that fidelity can be accomplished in real world settings. To this end, Horner et al. (2009) demonstrated in a randomized wait-list trial that SWPBIS training delivered by typical state level staff was functionally related to improved fidelity outcomes. Using the SET as an outcome metric, the study authors

documented that treatment schools were not implementing SWPBIS prior to intervention and state level trainers helped schools achieve scores higher than the 80% fidelity criteria at the end of study. This study provided causal evidence that SWPBIS can be implemented with fidelity in real-world rather than laboratory settings.

A repeated measure of SWPBIS fidelity. In terms of longitudinal measurement of implementation, SWPBIS researchers have developed the Team Implementation Checklist (TIC) (Sugai, Todd, et al., 2001) to monitor installation of program components during the early stages of implementation. The measure provides operational definitions of the critical components necessary for implementing SWPBIS with fidelity organized around themes of establishing commitment, establishing and maintaining a team, self-assessment, establishing school-wide expectations via prevention systems, classroom behavior support systems, and building capacity for function-based support. The tool is designed to be completed quarterly by school SWPBIS teams and used to develop action plans to support implementation activities. At this point, only limited psychometric information is available for this measure (i.e., Tobin, Vincent, Horner, Rossetto Dickey, & May, 2012). Given the design of the measure and repeated administration by schools within a school year, it provides an excellent opportunity to explore fidelity growth as construct and determine what factors predict this rate.

Predictors of SWPBIS fidelity growth. The TIC provides an opportunity to explore within year fidelity growth of the school system's intervention, and subsequently exploring predictors of this growth could provide documentation of variables associated with a dynamic systems process. To this end, several non-malleable and malleable factors that may be associated with fidelity growth are described.

Years implementing and other school contextual variables. Years implementing and school contextual variables represent non-malleable factors that may be associated with fidelity growth. Logically, schools that have been implementing SWPBIS for several years might be expected to have implemented more critical pieces of SWPBIS and subsequently have higher fidelity ratings as they approach what Fixsen and colleagues (2005) refer to as *full implementation*. In turn, their rate of change, or slope in linear growth modeling terms, would be less steep as they approach 100% implementation fidelity of critical program components. Conversely, schools in the *exploration* and *initial* stages of implementation (Fixsen et al., 2005) might logically have lower implementation ratings and steeper slopes as critical components are put into place. Schools contextual variables have been associated with student outcomes (Stone & Lane, 2003; Zvoch & Stevens, 2006) and represent possible non-malleable factors related to fidelity growth.

Sustainability factors. Sustainability factors represent malleable variables that may be associated with fidelity growth. Once a program has been implemented, there is no guarantee that the program will remain in place down the road. In fact, many innovations are not sustained after the initial implementation (e.g., Santangelo, 2009). Perhaps this is a product of changing priorities, or perhaps because of a lack of sensitivity and responsiveness between program providers and school staff (Vaughn, Klingner, & Hughes, 2000). Regardless, once a program is implemented resources for the program are often removed because, for example, critical components for the installation and initial implementation of a program are no longer needed and/or the original funding streams are no longer available. This provides greater stress on systems, as they are

required to maintain necessary drivers for consistent innovation use without external support. The documented relationship between implementation drivers and sustained use (e.g., Fixsen, Blase, Timbers, & Wolf, 2001) coalesces facilitators and barriers to maintaining programs beyond implementation stages into a model of sustainability.

SWPBIS provides an excellent opportunity for studying the consistent and sustained use (i.e., maintained implementation fidelity) of an educational innovation at scale. In this context, McIntosh, Filter, Bennett, Ryan and Sugai (2010) defined a model of sustainability in terms of *contextual fit*, *priority*, *effectiveness*, *efficiency*, and *continuous regeneration* and is based on numerous empirical examples documenting the relationship between implementation drivers and sustained use of SWPBIS and other evidence based practices. Further the authors contend

> The critical mechanism by which a practice sustains is fidelity of implementation… An effort in which school personnel continue to implement with low fidelity, or implement only noncritical features, does not meet the definition of sustainability. The target for a sustainability initiative is therefore the behavior of the school personnel, and targeting sustainability requires targeting the environment of the adults in the school (p. 10).

Contextual fit is the alignment of a practice to the needs of the school and district. *Priority* is the impetus to continue using a practice over time in spite of continually evolving contextual foci. *Effectiveness* is achieved through fidelity of implementation and positive results of the practice on student outcomes. *Efficiency* relates to the feasibility of using a practice to achieve desired outcomes. Finally, *continuous regeneration* relates to building capacity by using data to monitor implementation,

evaluating outcomes, and adapting as necessary. In sum, these pieces are hypothesized to be necessary components for sustaining the use of an evidenced based practice like SWPBIS and are supported by a research base documenting implementation facilitators and barriers. This model of sustainability and related empirical evidence are detailed in the following paragraphs.

Contextual fit is an involved process that requires stakeholders at various levels within an organization to view the importance of using a proposed practice. This requires various steps including preparing the argument of the practice's value, and mobilizing interest, consensus, and support from key stakeholders (Adelman & Taylor, 2003). Not only must individual stakeholders believe in the importance of the innovation, but school communities need to have a common understanding and appreciation for a systems intervention such as SWPBIS (Bambara, Nonnemacher, & Kern, 2009). In fact, contextual fit has been documented as both a facilitator when present and barrier when absent of effective SWPBIS implementation (Kincaid, Childs, Blase, & Wallace, 2007; Lohrmann, Forman, Martin, & Palmieri, 2008; McIntosh et al., in press). Further, contextual fit has been found to aid or impede the implementation of other types of educational programs including secondary transition programs (Benz, Lindstrom, Unruh, & Waintrup, 2004), school reform models (Berman & McLaughlin, 1976), and instructional innovation (Baker, Gersten, Dimino, & Griffiths, 2004; Datnow, Park, & Wohlstetter, 2007).

Fit is not enough to ensure that an innovation can be implemented and sustained effectively; the practice must be a priority. Priority is documented in several ways as a facilitator and/or barrier to effective implementation including: Staff commitment

(Berman & McLaughlin, 1976; Kincaid et al., 2007; Lohrmann et al., 2008; McIntosh et al., in press), district and school administrative support (Bambara et al., 2009; Benz et al., 2004; Berman & McLaughlin, 1976; Coffey & Horner, 2012; Datnow et al., 2007; Kincaid et al., 2007; Lohrmann et al., 2008; McIntosh et al., in press; Rohrbach, Graham, & Hansen, 1993; Santangelo, 2009), and funding (Kincaid et al., 2007). Communication has also been identified as an integral component to effective implementation and sustainability of SWPBIS (Coffey & Horner, 2012; Kincaid et al., 2007), and perhaps exemplifies priority via a manifestation of staff commitment and administrator support.

Effectiveness is a multidimensional construct that is made of several facets including (a) staff perceptions of a practice, (b) implementation fidelity and skill, and (c) positive impacts on student outcomes. In terms of staff perceptions, school personnel's attitudes and beliefs may change as a result of staff development, change in practices, and changes in student outcomes (Guskey, 1986). Perception and sustained use of a practice may also be related to teacher characteristics (Rohrbach et al., 1993; Sparks, 1988). Regardless of the mechanisms involved in altering perceptions, evidence has documented the relationship between effectiveness and sustaining SWPBIS (Kincaid et al., 2007; Lohrmann et al., 2008) and other educational innovations (Berman & McLaughlin, 1976; Rohrbach et al., 1993; Sparks, 1988). The effectiveness of using an innovation relates to the implementer's skill and knowledge of the practice (Baker et al., 2004; Berman & McLaughlin, 1976). Additionally, school personnel perceive effectiveness when data documents the program's effectiveness on improving desired outcomes in localized contexts (Han & Weiss, 2005).

Efficiency relates to the feasibility of using a practice to achieve desired outcomes. Adelmann and Taylor (2003) explained that clarifying feasibility can occur via explanation of the institutional functions necessary for adopting a practice, how changes will be accomplished, and formulating a long-range strategic plan for maintenance of the program. Another integral component is use of time. Past research has indicated that amount of time involvement influences judgments of interventions' accessibility (Witt, Martens, & Elliott, 1984) (a finding closely related to effectiveness as well). Further, Bambara and colleagues (2009) documented that lack of time for regular meetings and perceptions of additional burdens for program components were barriers to proper implementation of individualized positive behavior support programs. Efficiency, like the other constructs of sustainability, is multidimensional involving implementers' perceptions of time to feasibly implement programs and systemic adjustments to allow for adequate time to focus on the program. When deciding on adoption, schools must weigh what they will have to take away in order to add a program effectively as the program must efficiently interact with the other numerous school programs.

In a sense, the overall goal for understanding the facilitators and barriers to sustainability is to describe the drivers related to programs maintaining themselves after formal implementation processes have ended. Continuous regeneration is another multidimensional construct relating to program adaption and continuous capacity building. In this sense, it is a result of aligned implementation drivers and "use of data for decision making is the foundation for continuous regeneration" (McIntosh et al., 2010, p. 14).

Data based decision making research has focused on enhancing separate implementation drivers, as mentioned, and provides evidence documenting the relationship of integrated implementation drivers and continuous regeneration. Coupled with effective interventions for enhanced data-use practices (e.g., Todd et al., 2011), collection and use of data for decision making and monitoring of program outcomes has been documented as an integral piece of SWPBIS implementation and sustainability (Coffey & Horner, 2012; Kincaid et al., 2007). These examples highlight the integration of several implementation drivers including training, systems interventions, decision support data systems, and time to analyze data leading to sustained use SWPBIS. Studies of academic programs have also documented the influence of integrated implementation drivers and important outcomes. For example, collection and use of data was associated with sustained use of the practice (Baker et al., 2004), and continued focus on a data driven reform process three years following implementation was associated with enhanced academic outcomes (Slavin et al., 2010). These examples highlight implementation drivers (decision support data systems and time) interacting to facilitate valued results. Interestingly, concurrent across several of these studies is the use of a teaming structure, which aligns with Odom's (2009) model of implementation describing enhanced professional development as including a teaming, systemic support through ongoing professional development and support.

The use of data can aid continuous regeneration via program maintenance adaptation. In fact, maintenance might be a function of success during the implementation phase and continued focus on outcomes to monitor and adapt program delivery (Han & Weiss, 2005, p. 676). Coburn (2003, p. 7) refers to this as a "shift in

ownership" with the authority moving from external reform agents such as researchers to district and school personnel. While transitioning authority is important, empirical evidence has shown the sustainability of behavioral programs such as SWPBIS was related to ongoing access to professional development and support (Bambara et al., 2009; Kincaid et al., 2007; Mathews, McIntosh, Frank, & May, in press). Seemingly, district and school leaders must make decisions to allocate personnel, training, and time to maintain programs once external supports are no longer available. Further, these decisions can aid to a sustained focus on data and program outcomes to ensure that adaptations lead to continued success.

In response to literature highlighting the facilitators and barriers of sustained implementation of innovations, a measure was developed to assess school staff's perceptions of these factors when implementing SWPBIS. The School-wide Universal Behavior Sustainability Index: School Teams (SUBSIST) (McIntosh et al., 2009) is a contextual measure of implementation relating to perceptions of sustainability factors aimed at determining how conducive a school environment is to sustaining use of SWPBIS. Evidence has documented the SUBSIST's content validity, reliability, and concurrent validity with fidelity of implementation (McIntosh et al., 2011), and the predictive validity of the factor structure and sustained implementation (Mcintosh et al., 2013). Additionally, invariance tests of the factor structure have documented measurement invariance across schools at similar phases of implementation (Mercer, McIntosh, Strickland-Cohen, & Horner, Manuscript submitted). The factors labeled *school priority*, *team use of data*, *district priority*, and *capacity building* related directly to organizational drivers' influence on sustaining use of SWPBIS.

Considering the tool and previous findings, SUBSIST factor scores potentially could be used to understand how perceptions of schools' ability to sustain SWPBIS influence the implementation process. Programs in the *installation* and *initial implementation* stages could benefit from continuous feedback to ensure integral program and systemic pieces are in place to promote *full implementation* and sustainability after the stages of implementation are complete. In educational contexts, providing feedback to schools and districts in real-time to ensure effective implementation and sustainability would help ensure that the intents of school reforms and improvements are realized.

Study Aims

One main goal of this study is to document fidelity growth of SWPBIS to describe varying rates of implementation between schools. Fidelity growth will be defined as the growth in the number of implementation tasks that have been fully and partially completed. Specifically, this study will model the linear growth of within-year rate of SWPBIS fidelity growth as measured by the TIC. Another main focus of this study is to determine if school and district variables predict implementation fidelity growth of an evidenced based program. Specifically, do nonmalleable and malleable contextual variables such as relative socio-economic status, the number of years implementing, and SUBSIST factor scores predict rate of implementation? Findings will potentially provide psychometric evidence for a SWPBIS fidelity growth metric. Further, these results could benefit schools implementing SWPBIS by providing a specific link between the implementation process and systems' variables related to the practice.

As with many scientific pursuits, the challenge of the present study is having a large enough sample to link the desired variables. Given the limitation that the only a

small number of schools who completed the prerequisite number of measures per year to model linear growth, alternative methods need to be explored. That is, only a small number of schools completed the TIC three or more times per year and made their data available to researchers. Subsequently, this limitation provides the opportunity to demonstrate the use Bayesian methods to model linear growth and to serially update growth estimates over several samples by using results from each year's data collection to inform subsequent year's statistical estimates, a method known as Bayesian Updating.

CHAPTER III

METHODS

The goal of this study was to demonstrate the use of Bayesian estimation and Bayesian Updating (BU) techniques in a hierarchical growth model and compare estimates to a maximum likelihood (ML) approach. In this respect, a thorough procedure for conducting a Bayesian analysis over multiple waves of data collection is outlined and demonstrated. The goal of which is to show the utility of these techniques for educational researchers. Specifically, the study demonstrated the use of Bayesian methods for modeling the fidelity growth of a school-wide behavior program across multiple samples from a similar population, and determined if the number of years implementing, contextual variables, and program sustainability factors predict this growth.

Design and Samples

This study used secondary data collected by a SWPBIS technical assistance center (i.e., the OSEP Technical Assistance Center on Effective Schoolwide Interventions: Positive Behavioral Interventions and Supports, www.pbis.org) and two previous studies (i.e., Mcintosh et al., 2013; Mercer et al., Manuscript submitted). School-level data collected over five school years ranging from 2008-09 to 2012-13 was used. Each year's data were considered a separate sample in accordance with the Bayesian method employed. In sum, five samples comprised the entirety of the data for this study.

Participants

Participants included SWPBIS school teams that are comprised of building leaders and district coaches from schools throughout the United States. As the analyses

for this study were conducted at the school-level, descriptions of schools in each sample are detailed in Table 1. School characteristics were collected by the Institute for Education Science during the 2009-10 school year (National Center for Education Statistics, 2011). Sample sizes ranged from 13 schools in 2008-09 to 85 schools in 2012-13. The states represented in all of the samples were California, Idaho, Minnesota, Missouri, Oregon, and Wisconsin. Minnesota was the most represented state with 157 participating schools across the five samples. Oregon was the least represented state with three participating schools across the five samples all of which were from the 2008-09 school year. While the majority of schools for the 2008-09 sample were located in towns and rural communities, all types of communities were represented across samples. The majority of schools sampled were elementary schools and Title I eligible[3].

The schools' student bodies ranged in size and composition. For example, the smallest school across all samples had an enrollment size of 13 students (from the 2012-13 sample) while the largest school had an enrollment size of 2,112 students (from the 2010-11 sample). The percent of students eligible for free and reduced price lunch ranged between samples, and was about 42% across all samples. Males comprised a little over 51% for each sample and across all samples. The majority of students in each school were White at over 75% across samples. Latino students were the second highest represented racial/ethnic group with slightly more than 6% across all samples.

[3] Title I eligibility is defined as having at least 40% of students from low income households (Office of Student Achievement and School Accountability Programs, 2011).

Table 1

Demographic Characteristics for Schools in Each Sample, A Subset of the 2010-11 Sample with Sustainability Data, and for a Pooled Sample with Schools from All Years

	2008-09 (*j = 13)*	2009-10 (*j* = 61)	2010-11 (*j* = 53)	2010-11**** (*j* = 10)	2011-12 (*j* = 64)	2012-13 (*j* = 85)	Total (*j* = 276)
# states represented	2	4	5	3	4	4	6
Locality							
City	1	18	16	5	7	27	69
Suburb	1	7	19	3	36	29	92
Town	6	13	8	1	10	14	51
Rural	5	23	10	1	11	15	64
# full time employees*	26.0	30.3	33.1	30.5	34.0	33.7	32.8
School level							
Primary	7	39	44	10	46	55	191
Middle	5	14	5	0	9	20	53
High	1	7	4	0	6	7	25
Title I eligible							
Yes	9	41	37	7	39	48	174
No	4	20	16	3	25	37	102
Student body characteristics**							
# students	505.6	452.5	588.6	496.8	562.0	631.4	561.7
j schools w/ <= 500 students	7	41	31	7	27	35	141
j schools w/ > 500 students	6	20	22	3	37	50	135
% FRL eligible***	61.6	49.1	41.3	39.4	37.1	36.6	41.8
% male	51.3	52.7	51.4	51.7	51.2	51.2	51.5
% female	48.7	47.3	48.6	48.3	48.8	48.8	48.5
% American Indian/ Alaskan Native	0.7	0.6	0.4	0.7	0.5	0.3	0.5
% Asian/ Pacific Islander	0.9	1.1	3.9	8.8	5.6	4.3	3.1
% Black	1.4	2.3	4.5	10.9	5.6	2.7	3.3
% Hawaiian/ Pacific Islander	0.1	0.0	0.0	0.0	0.0	0.0	0.0
% Latino	3.2	3.1	5.7	7.8	8.1	9.3	6.1
% two or more races	0.5	0.7	2.0	3.1	2.2	1.8	1.8
% White	89.5	89.8	76.6	67	71.1	68.2	75.1

Note. *# of full time employees is arithmetic mean for all schools in the sample. **# of students is the arithmetic mean for all schools in the sample, and student characteristics by percentage is the median percent across all schools in sample. ***Free and reduced priced lunch eligible students. ****SUBSIST sample for the 2010-11 school year (a subset of the 2010-11 sample).

Measures

Team Implementation Checklist (TIC). The TIC (Sugai et al., 2002, 2009; Sugai et al., 2011) measured fidelity of SWPBIS during the initial program implementation. As a fidelity measure, the TIC measured adherence to the critical features of SWPBIS. While its use varies widely (Tobin, 2006), it was designed to be completed quarterly throughout a school year allowing the examining of fidelity change over the course of a year. SWPBIS teams and district coaches self-administered the measures during SWPBIS team meetings periodically throughout the school year. The TIC was designed to be used during the first few years of implementation to create and modify a SWPBIS implementation action plan. Once a school has achieved an 80% completion rate on three occasions, alternative evaluative tools are recommended. No psychometric information is available for version 2.1 and 3.1. Research on TIC version 3.0 documented a high internal consistency (Cronbach $\alpha = .91$), and concurrent validity ($r = .59$) with the Benchmarks of Quality, a SWPBIS fidelity measure designed to measure the implementation of critical SWPBIS features and evaluate their effectiveness (Tobin et al., 2012).

Schools completed three different versions (2.2 (Sugai et al., 2002), 3.0 (Sugai et al., 2009), and 3.1 (Sugai et al., 2011)) of the TIC between 2008-13. Further, the study involved data collected across samples. Borrowing techniques from integrative data analysis, measurement harmonization techniques were employed to create an equivalent measure across samples (Hussong, Curran, & Bauer, 2013). To this end, scale scores were created from the 17 items consistent across all versions. The 17 items represented essential steps for initial and on-going implementation of SWPBIS and were reported on

a three-point ordinal scale (*not yet started*, *in progress*, and *achieved*). Points were assigned to each position on the scale with zero indicating *not yet started*, one for *in progress* responses, and two for *achieved*. Table 2 depicts the wording of items for each of the three versions, and highlights the consistent items across versions. The wording changed between version 3.0 and 3.1, but the critical features of SWPBIS are still represented. Specifically, items on version 3.1 exemplify the items' intent with specific behavioral definitions. The wording of item 17 on version 2.2 and item 22 on version 3.0 deviates more than other items that were consistent across versions. This item was considered consistent across versions because the item relates to programmatic features involving the use of function based support. Scale scores were created by adding the total score across all 17 consistent items with a highest score being 34. If a school indicated that an item was (a) complete, a two was recorded; (b) in progress, a one was recorded; and (c) not yet started, a zero was recorded. TIC version 3.1 is available in Appendix A.

School-wide Universal Behavior Sustainability Index: School Teams (SUBSIST). The SUBSIST (McIntosh et al., 2009) is a survey measure completed by SWPBIS team members and is reprinted in Appendix D. It contains items completed using a four point Likert-type scale (ranging from *not true* to *very true*) relating to facilitators and barriers for sustaining SWPBIS. An initial study documented the content validity of the questions, and the reliability in terms of internal consistency of the subscales (alpha coefficients ranged from .77 to .94), test-retest reliability (r = .96), and inter-rater reliability (r = .95) (McIntosh et al., 2011). Concurrent validity between the

Table 2

Team Implementation Checklist Start-Up Activity Items for Various Versions

Domain	2.2	3.0	3.1	Consistent*
Establish commitment	1. Administrator's support & active involvement.	1. Administrator's support & active involvement.	1. Administrator's Support & Active Involvement • Admin attends PBIS meetings 80 % of time • Admin defines social behavior as one of the top three goals for the school • Admin actively participates in PBIS training	yes
	2. Faculty/Staff support (One of top 3 goals, 80% of faculty document support, 3 year timeline).	2. Faculty/Staff support (One of top 3 goals, 80% of faculty document support, 3 year timeline).	2. Faculty/Staff Support • 80% of faculty document support that school climate/discipline is one of top three school improvement goals • Admin/faculty commit to PBIS for at least 3 years	yes
Establish and maintain team	3. Team established (representative).	3. Team established (representative).	3. Team Established (Representative) • Includes grade level teachers, specialists, paraprofessionals, parents, special educators, counselors. • Team has established clear mission/purpose	yes
	4. Team has regular meeting schedule, effective operating procedures.	4. Team has regular meeting schedule, effective operating procedures.	4. Team has regular meeting schedule, effective operating procedures • Agenda and meeting minutes are used • Team decisions are identified, and action plan developed	yes
	5. Audit is completed for efficient integration of team with other teams/initiatives addressing behaviour support.	5. Audit is completed for efficient integration of team with other teams/initiatives addressing behavior support.	5. Audit is completed for efficient integration of team with other teams/initiatives addressing behavior support • Team has completed the "Working Smarter" matrix	yes
Self-assessment	6. Team/faculty completes PBS self-assessment survey.	6. Team/faculty completes the Team Checklist or Benchmarks of Quality self-assessment	6. Team completes self-assessment of current PBIS practices being used in the school • The staff completes the TIC (progress monitoring), BoQ (annual assessment) or SET.	yes
	7. Team summarizes existing school discipline data.	7. Team summarizes existing school discipline data.	7. Team summarizes existing school discipline data • The team uses office discipline referral data (ODR), attendance, & other behavioral data for decision making.	yes
	8. Strengths, areas of immediate focus & action plan are identified.	8. Team uses self-assessment information to build implementation action plan.	8. Team uses self-assessment information to build implementation Action Plan (areas of immediate focus) • The team uses the Action Plan to guide PBIS implementation.	yes
Establish school-wide expectations**	9. 3-5 school-wide behavior expectations are defined.	9. 3-5 school-wide behavior expectations are defined.	9. 3-5 school-wide behavior expectations are defined and posted in all areas of building • 3-5 positively and clearly stated expectations are defined. • The expectations are posted in public areas of the school.	yes

Table 2 (continued)

Domain	2.2	3.0	3.1	Consistent*
	10. School-wide teaching matrix developed.	10. School-wide teaching matrix developed.	10. School-wide teaching matrix developed • Teaching matrix used to define how school-wide expectations apply to specific school locations. • Teaching matrix distributed to all staff.	yes
	11. Teaching plans for school-wide expectations are developed.	11. Teaching plans for school-wide expectations are developed.	11. Teaching plans for school-wide expectations are developed• Lesson plans developed for teaching school-wide expectations at key locations throughout the school.• Faculty is involved in development of lesson plans.	yes
	12. School-wide behaviour expectations taught directly & formally.	12. School-wide behavioral expectations taught directly & formally.	12. . School-wide behavioral expectations taught directly & formally • Schedule/plans for teaching the staff the lessons plans for students are developed • Staff and students know the defined expectations. • School-wide expectations taught to all students • Plan developed for teaching expectations to students to who enter the school mid-year.	yes
	13. System in place to acknowledge/reward school-wide expectations.	13. System in place to acknowledge/reward school-wide expectations.	13. System in place to acknowledge/reward school-wide expectations • Reward systems are used to acknowledge school-wide behavioral expectations. • Ratio of reinforcements to corrections is high (4:1). • Students and staff know about the acknowledgement system & students are receiving positive acknowledgements.	yes
	14. Clearly defined & consistent consequences and procedures for undesirable behaviours are developed.	14. Clearly defined & consistent consequences and procedures for undesirable behaviors are developed.	14. Clearly defined & consistent consequences and procedures for undesirable behaviors are developed • Major & minor problem behaviors are all clearly defined. • Clearly defined and consistent consequences and procedures for undesirable behaviors are developed and used. • Procedures define an array of appropriate responses to minor (classroom managed behaviors). • Procedures define an array of appropriate responses to major (office managed) behaviors.	yes
Classroom behavior support systems		15. Team has completed a school-wide classroom systems summary	15. chool has completed a school-wide classroom systems summary • The teaching staff has completed a classroom assessment (Examples: SAS Classroom Survey, Classroom Systems Survey, etc.)	no
		16. Action plan in place to address any classroom systems identified as a high priority for change.	16. Action plan in place to address any classroom systems identified as a high priority for change • Results of the assessment are used to plan staff professional development and support.	no
		17. Data system in place to monitor office discipline referral rates that come from classrooms.	17. Data system in place to monitor office discipline referral rates that come from classrooms • School has a way to review ODR data from classrooms to use in data based decision making.	no

Table 2 (continued)

Domain	2.2	3.0	3.1	Consistent*
Establish information system***	15. Discipline data are gathered, summarized, & reported.	18. Discipline data are gathered, summarized, & reported at least quarterly to whole faculty.	18. Discipline data are gathered, summarized, & reported at least quarterly to whole faculty• Data collection is easy, efficient & relevant for decision-making• ODR data entered at least weekly (min).• Office referral form lists a) student/grade, b) date/time, c) referring staff, d) problem behavior, e) location, f) persons involved, g) probable motivation, h) consequences and i) administrative decision.• ODR data are available by frequency, location, time, type of problem behavior, motivation and student.• ODR data summary shared with faculty at least monthly (min).	yes
		19. Discipline data are available to the Team at least monthly in a form and depth needed for problem solving.	19. Discipline data are available to the Team regularly (at least monthly) in a form and depth needed for problem solving • Team is able to use the data for decision making, problem solving, action planning and evaluation. • Precision problem statements are used for problem solving.	no
Build capacity for function-based support	16. Personnel with behaviour expertise are identified & involved.	20. Personnel with behavioral expertise are identified & involved.	20. Personnel with behavioral expertise are identified & involved • Personnel are able to provide behavior expertise for students needing Tier II and Tier III support.	yes
		21. At least one staff member of the school is able to conduct simple functional behavioral assessments.	21. At least one staff member of the school is able to conduct simple functional behavioral assessments • At least one staff member can conduct simple behavioral assessments and work with a team in developing behavior support plans for individual students	no
	17. Plan developed to identify and establish systems for teacher support, functional behaviour assessment & support plan development & implementation.	22. Intensive, individual student support team structure in place to use function-based supports.	22. Intensive, individual student support team structure in place to use function-based supports • A team exists that focuses on intensive individualized supports for students needing Tier III supports. • The team uses function-based supports to develop, monitor and evaluate behavioral plans. • The team delivering Tier III has a data system that allows on-going monitoring of the fidelity and outcomes of individual behavior support plans.	yes
On-going activity monitoring****	1. PBS team has met at least monthly.			no

Table 2 (continued)

Domain	2.2	3.0	3.1	Consistent*
	2. PBS team has given status report to faculty at least monthly.			no
	3. Activities for PBS action plan implemented.			no
	4. Accuracy of implementation of PBS action plan assessed.			no
	5. Effectiveness of PBS action plan implementation assessed.			no
	6. PBS data analyzed.			no

Note. All items had a similar response scale: 'not started', 'in progress, and achieved'. For version 3.1, the response option for 'not started' was labeled 'not yet started'. The 17 consistent items were used to form a fidelity scale where observed scores were calculated by (a) coding 0 for 'not started', 1 for 'in progress', and 2 for 'achieved', and (b) summing response scores across all 17 items. *Consistent is defined as referring to the same SWPBIS program component. **Labeled Establish school-wide expectations: Prevention systems for version 3.0 and 3.1. ***This domain did not exist on version 3.1, and items 18 and 19 were included in the classroom behavior support sysems domain. ****This domain did not exist on versions 3.0 and 3.1.

SUBSIST and the School-wide Evaluation Tool (Horner et al., 2004), an evaluative tool for documenting implementation fidelity of critical SWPBIS features, was moderate (r = .68) demonstrating that the SUBSIST measured a similar but unique construct (McIntosh et al., 2011, p. 213). Factor analysis revealed a four factor structure for the SUBSIST with two school factors (labeled *school priority* and *team use of data*) and two district factors (*district priority* and *capacity building*); *team use of data* and *capacity building* were significant predictors of sustained SWPBIS fidelity (Mcintosh et al., 2013). Finally, factor invariance has been documented between schools with varying number of years implementing SWPBIS (*0 to 1 year*, *2 to 4 years*, and *5 or more years*) (Mercer et al., Manuscript submitted). SUBSIST factor scores were estimated using a confirmatory factor analytic model in accordance with McIntosh et al. (2013). As the measure was completed by school level personnel, district factors were calculated by taking the mean of district factor scores of all personnel in that district who completed the SUBSIST.

Factor scores were estimated using a large sample of schools that completed the measure using the online database and multiple imputation was used to account for missing responses to survey items. 860 schools using the online SWPBIS database entered SUBSIST data for the 2010-11 and 2012-13 school years. A three phase multiple imputation technique was used to handle missing survey response data as a result of missing and 'I don't know' responses based on recommendations from Enders (2010) using a similar technique employed by McIntosh and colleagues (2013). The three phase technique incorporated an imputation, analysis, and pooling phase. For the imputation phase, ten data sets were estimated via multiple imputation using the 'mi' package (Su, Gelman, Hill, & Yajima, 2011) for the R statistical program (R Core Team, 2012). For

the analysis phase and in accordance with McIntosh and colleague's (2013, pp. 301-302) model, 10 separate factor analyses were conducted using the lavaan (Rosseel, 2012) package with a variance adjusted weighted least squares (WLSMV) estimator, and items specified as ordered categorical. Lavaan was utilized because of its similarities to Mplus (L. K. Muthén & Muthén, 1998-2012). Factor scores for each of the original 860 schools were estimated using the predict function for each of the 10 estimated models. For the pooling phase, the arithmetic means for each factor score of each school were used as covariates in later analyses.

School context variables. School context variables included years implementing SWPBIS, locality, school size, and relative socio-economic status. The number of years implementing was obtained from an online SWPBIS database. Locality, school size, and relative socio-economic status were contextual variables for each school that were either gathered directly or by converting data collected by the National Center for Education Statistics (NCES) during their 2009-10 school census (National Center for Education Statistics, 2011).

- The number of years implementing SWPBIS was defined as the number of complete school years the school has been actively submitting data into an online database before the measurement occasion. This variable was marked zero for schools in their first year of implementation, a one designated schools that have completed one full year of implementation and were in their second year, a two designated schools that have completed two full years, and so on.
- Locality was truncated from twelve original categories of school's location relative to population ranging from city-large to rural-remote, to four location categories defined

as city, suburb, town, and rural. In reference to the original codes from NCES (2011), city was defined as territory inside a urban area and inside a principal city, suburb as territory inside an urban area and outside a principal city, town as inside an urban cluster and outside an urbanized area, and rural as at least 5 miles from an urbanized area and 2.5 miles from an urban cluster.

- School size was obtained directly from the NCES (2011) data, and was defined in terms of the number of students attending the school during the 2009-10 data collection. This variable was then converted to a binary indicator with a zero representing schools with less than or equal to 500 students, and a one for schools with more than 500 students.
- Relative socio-economic status was created via a proxy calculated by dividing the number of students eligible for free and reduced priced lunch and the total number of students in the school. This may not be the most valid proxy for relative socio-economic status of a school (Harwell & LeBeau, 2010), so interpretations will be related to its definition. This created a variable labeled percentage of students eligible for free and reduced price lunch.

Procedure

For this exploratory study, schools were selected for a particular year's sample if they completed the requisite measure(s) during that year. Additionally, schools included in the 2010-11 and 2012-13 samples were recruited via specific mechanisms: For the 2010-11 sample, schools were recruited via invitations sent to school and district personnel from state SWPBIS coordinators (Mcintosh et al., 2013); for the 2012-13 sample, state SWPBIS teams recruited schools during training events and via emailed

invitations (Mercer et al., Manuscript submitted). School personnel completed the SUBSIST via an online survey tool. School SWPBIS teams completed the TIC during team meetings and entered the results into an online database.

For each sample, schools were included from the database if for that particular school year they (a) completed three or more TICs, and (b) were in years zero to three of program implementation. The criterion three or more TICs per school year was chosen to facilitate the multilevel growth modeling technique utilized for this methods demonstration. Given this restriction relating to the TIC, schools included in the final samples for each year may not generalize to the greater population of schools using SWPBIS. As the main focus of this study was to demonstrate Bayesian techniques for multilevel growth modeling, this restriction was less problematic than if the main goal was to generalize to all schools implementing SWPBIS. Nevertheless, limitations will be discussed. The reason for including those schools in the initial years of implementation related directly to the TIC measure's intended use of documenting implementation fidelity as schools progress towards full SWPBIS program fidelity during the first few years of implementation.

Analysis

Preliminary analysis. Descriptive analyses were conducted to describe the schools in each sample. Schools were characterized in terms of their years implementing SWPBIS, location in the U.S., grade levels, enrollment size, and region (i.e., suburban, urban, rural). Descriptive statistics were provided for the outcome variable, TIC scores, and covariates included in the models.

Missing data. Given the inclusion procedure described, the main missing data concern for this study was in the form of missing survey responses on the SUBSIST measure. Considerations were made for missing survey responses as a result of missing and 'I don't know' responses. The specific method employed is outlined in the section on the SUBSIST measure.

Analytic approach. Models were estimated using a linear growth model utilizing fully Bayesian estimation techniques (Gelman et al., 2004; Gelman & Hill, 2007; Gill, 2002; Kuiper et al., 2013). Maximum likelihood (ML) models were estimated to make comparisons between techniques and discuss the likelihood distributions' influence on Bayesian posterior estimates. Several options exist for modeling growth including auto-correlation techniques (e.g., Wang & Daniels, 2013), latent class growth analysis (Kline, 2011), and multilevel/hierarchical regression models (Raudenbush & Bryk, 2002). For this analysis, multilevel regression models were used as they accommodate an unequal number of measurement occasions and unequal spacing between time points for each case (Raudenbush & Bryk, 2002). Curvilinear models were not considered as schools included in the samples typically completed the TIC on three occasions. I used the Just Another Gibbs Sampler (JAGS) package (Plummer, 2003) for the R statistical program (R Core Team, 2012) to estimate the models. Considering the complexity of incorporating all these pieces, model assumptions were explored as they were introduced into analyses by (a) model specification and likelihood distributions, (b) Markov Chain Monte Carlo sampling, and (c) specification of prior distributional forms. Finally, model fit was assessed to determine the adequacy of estimated models for documenting the phenomenon being studied. Considerations for model assumptions and model fit for all

research questions are outlined in the following sections, and then specific considerations for each research question are addressed.

For the coding of time, time was centered at the first measurement occasion for each school in each sample. Several options were considered, including centering at the end of each school's year. The data available created limitations in this respect as (a) schools did not uniformly complete the TIC at the beginning and/or end of each school year, and (b) accurate data on the beginning and ending date of each school's year is not available. These limitations subsequently impeded the creation of a substantively meaningful intercept. That is, creating a parameter for the beginning or end of the school year would be potentially biased, as these time points could not be accurately defined. This decision on coding time and the use of multilevel modeling enabled estimation of linear change in fidelity over time for each school in each sample. For the chosen procedure, time intervals were coded in terms of months from the initial time point, and each month was defined as thirty days. For each measurement occasion, the data from the online database included a time stamp of month, day and year enabling the calculation of days elapsed between measurement occasions

Model specification and Maximum Likelihood (ML) estimation. Like all analyses, use of Bayesian estimation requires special attention to the form of the outcome variable. Additionally, the JAGS software requires that models be specified in the form of the outcome variable. For this study, the outcome variable was the score on the TIC and followed the form of a normal distribution. Models were specified based on sample code outlined in Gelman and Hill (2007) that employ a multivariate normal likelihood distribution. Sample code is provided in Appendices C, D, and E.

Markov Chain Monte Carlo techniques. A fully Bayesian approach incorporates both the likelihood estimates and prior distributions using a Markov Chain Monte Carlo (MCMC) sampling technique entitled Gibbs sampling. The sampling requires both statistical and visual checks to determine that the separate chains adequately mixed and estimates are stable. The Gelman and Rubin convergence diagnostic, $\hat{R}$, was used as a numerical check to ensure chains have adequately mixed, and values less than or equal to 1.1 for each parameter generally indicate adequate mixture of Markov Chains (Gelman & Hill, 2007, p. 358). Additional visual checks of the chains' trace plots were used to confirm that the chains have adequately mixed. Documentation is provided if chains did not mix adequately. The number of simulations, the number of chains, and thinning rate (i.e., only saving every n^{th} iteration) were reported.

Specifying prior distributions. Using a fully Bayesian approach requires specification of prior distributions for every parameter in the model. That is, every parameter has hyper-parameters to specify the shape and size of the prior distribution for that parameter. Further, specification of prior distributions can range from completely unknown to very specific. Unknown priors are labeled *uninformative* and can be in various forms including uniform or conjugate distributions. Specific priors are labeled *informative* and hyper-parameters are based on anything from expert opinion or results from previous samples of data. Bayesian updating (BU) is a specific use of *informative priors* where posterior distributions from previous data samples are incorporated as prior distributions in subsequent samples. For this study, all use of *informative* priors was based on BU techniques using posterior results from previous samples as prior distributions in subsequent samples' models.

Uninformative prior distributions. *Uninformative* prior distributions were specified as part of analyses for every research question. *Uninformative* priors for all model parameters can take on many forms including uniform and conjugate distributions[4]. Uniform priors assume that the parameter values are completely unknown within a wide range of a bounded integral that integrates to one (Gill, 2002, pp. 120-121). Another option is to use conjugate priors that have a specified shape and ease calculation of posterior distributions (Gill, 2002, p. 115). Conjugate prior distributions, may be vague and cover a wide a range, a condition referred to as *diffuse*, and subsequently still considered *uninformative*. Special attention must be paid to the plausibility of results obtained when using conjugate priors because they can lead to inaccurate posterior estimates (Natarajan & McCulloch, 1998). Further, choice of conjugate distributions is based on the form of the likelihood distribution and introduces an assumption about the form of the prior distribution.

Informative Prior Distributions: Applying Bayesian Updating. As this study will apply the use of BU, posterior distributions of fixed effects from previous sample results were incorporated as *informative* priors of subsequent sample results. BU methods have been shown to improve model fit and robustness in hierarchical models (Yu & Abdel-Aty, 2013). For each model that incorporates prior distributions, three options were used to form posterior distribution estimates that incorporate a spectrum of specificity for the prior ranging from *somewhat informative'* to *informative* to *very informative*.

[4] Another option for specifying uninformative priors is to use Jeffrey's priors, which avoid subjective choice for specifying form and size of distributions. The use of Jeffrey's Priors can be very useful in single parameter models, but will not be explored because of the complexities of expanding the approach in multivariate models (Gill, 2002, p. 125).

The range of specificity for informative prior distributions involved varying hyper-parameters for fixed effects. Considering that the fixed effects followed a normal distribution, the hyper-parameters of the prior were specified with (a) the mean equal to the posterior parameter mean from the previous sample, and (b) the variance equal to a range of possibilities from 'somewhat informative' to 'very informative'. I defined 'very informative' variance hyper-parameters as the variance hyper-parameters for the previous sample's posterior distribution, and 'informative' and 'somewhat informative' as iteratively larger to encompass varying degrees of possibility. Specifically, (a) 'very informative' hyper-parameters were defined as the mean and standard deviation of the previous samples' posterior distributions, (b) 'informative' hyper-parameters were defined as the mean and two times the standard deviation of the previous sample's results, and (c) 'somewhat informative were the mean three times the standard deviation of the previous sample's poster distribution. In turn, the degree that the range of specificity for the variance hyper-parameters impacted posterior estimates of the fixed effects was described. As this study mainly focuses on the fixed effects, informative prior distributions were only used for the fixed effects. Because many research questions focus exclusively on fixed effect parameters, informative priors for random effects were excluded from this demonstration. Further, specifying informative priors for random effects involves using less common distributions such as gamma and Wishart distributions, which heightens the potential for model specification errors. Random effects were specified using uninformative techniques including uniform and conjugate prior distributional forms. Considering this procedure, close attention to model fit was necessary, and the methods are described in the following sections.

Model fit. When using complex statistical models and deciding on model adequacy, it seems appropriate to remember that model fit "statistics do not indicate whether the results are theoretically meaningful" (Kline, 2011, p. 193). For this study, I followed the general recommendation from Morey, Romeijn and Rounder (2013) and evaluated whether the statistical inferences supported by the model reasonably support real-world inferences. This involved global and local sensitivity analysis, robustness evaluation, and posterior predictive checking using recommendations from Gill (2002).

Global sensitivity analysis involved varying the widest possible range of assumptions, whereas local sensitivity analysis involved modest variations in assumptions. Specifically, for the global sensitivity analysis the (a) forms for the prior distributions were allowed to vary, as well as (b) other declared model assumptions. The local sensitivity analysis involved making modest changes to the prior parameterization by varying hyper-parameters, but keeping the form (e.g., normal, uniform, etc.) constant.

Robustness evaluation can be divided into two camps. The first camp, classical robustness, involves evaluating the model's sensitivity to influential outliers. This may more appropriately be called resistance to influential outliers as robust is an adjective for a model that has low sensitivity to violations of model assumptions (Gill, 2002, p. 171). Bayesian robustness refers to assessing the posterior sensitivity to substantively reasonable changes to model assumptions. A non-robust model would be one where reasonable changes to the prior and likelihood functions results in large changes in the range of posterior quantities of interest. Similar to global and local sensitivity analysis, a global and local robustness evaluation explored whether varying the prior distribution form and hyper-parameters impacted posterior distributions of interest, respectively.

Model selection. As model selection can be imprecise, several numerical indexes and tests were considered. Models were compared using AIC, BIC, DIC, and the approximate Bayes Factor described in Equation 4. Converging evidence across indices provided evidence for generalizability of one model over others. Definitions and considerations for each of these criteria are provided in the literature review.

Research question 1: Posterior distributions of fidelity growth after one sample. A simple growth model with varying intercepts and slopes was specified to answer the first research question:

$$\begin{aligned} &\text{Level 1:} \quad y_{tj} = \pi_{0j} + \pi_{1j} * time_t + e_{tj} \\ &\text{Level 2:} \quad \begin{aligned} \pi_{0j} &= \beta_{00} + u_{0j} \\ \pi_{1j} &= \beta_{10} + u_{1j}, \end{aligned} \end{aligned} \tag{5}$$

where y_{tj} equals the TIC score for school j at time t, π_{0j} is the parameter for the initial status that varies between schools, π_{1j} is the parameter for rate of change that varies between schools, $time_t$ is the time variable for each observation, and e_{tj} is the random effect for school j at time t. For the level two equations, β_{00} and β_{10} represent the average initial status and rate of change across all schools, respectively. Finally, u_{0j} and u_{1j} represent the growth parameter random effects for each school. The between school level variance/covariance was specified according to Tao matrix specified in equation 6 and will be assumed to be independent from the level one residual, e_{tj}:

$$T = \begin{bmatrix} \text{var}(u_{0j}) & - \\ \text{cov}(u_{0j}, u_{1j}) & \text{var}(u_{1j}) \end{bmatrix}. \tag{6}$$

For the main effects, β_{00} and β_{10}, both uniform and conjugate *uninformative* distributions were explored. For uniform distributions, the range of values was specified

to match the range of plausible values on the outcome measure. Specifically, the outcome measure was on the scale of 0 to 34, and the uniform distribution was specified to match this range. The conjugate distributions for the main effects were normally distributed to align with the multivariate normal distribution of the likelihood surface.

The *uninformative* prior distributions for the variance and covariance components of the model were assigned different forms including uniform, and various conjugate options to align with the multivariate normal likelihood distribution. For the level one error variance, gamma and uniform distributions were explored. For level two variance, models using uniform, gamma, and Wishart distributions were fit. Comparison between models was made using model selection techniques outlined.

Research question 2: Posterior distributions of implementation growth estimated using informative prior distributions. Models for the second research question were specified nearly identically to models for the first research question with the only change being in the form and size of the prior distribution for the fixed effects, π_{0j} and π_{1j}. *Informative* priors were used to update the form and size of the prior distributions for the analyses of samples collected in years two through five. This procedure permitted four applications of BU demonstrating the extent to which informative priors influenced implementation growth parameters over multiple samples. Informative priors for the fixed effects were specified according to previously outlined procedures where several models were estimated using various specifications of prior distributions. Specifically, prior distributions for the fixed effects were specified as distributed normally with a mean equal to the previous sample's mean and standard deviation ranging from 'somewhat specific' to 'very specific'. As mentioned previously,

the prior distribution standard deviations for fixed effects were based on the posterior distribution standard deviation from the previous sample with 'very specific' referring to the exact posterior standard deviation, 'specific' being twice the posterior standard deviation, and 'somewhat specific' being three times posterior standard deviation. Prior distributions for the random effects were specified with *uninformative* uniform and conjugate distributions, as this demonstration is limited to informative priors for fixed effects. Because many models were fit to investigate this range of form and specificity for prior distributions, model fit diagnostics were compared to judge which model was most adequate.

Research question 3: Posterior distributions of fixed effects predicting fidelity growth estimated with informative prior distributions. For the third research question, three separate analyses were conducted to demonstrate how to include informative prior distributions of fixed effects into a multilevel linear growth model using Bayesian updating. Specifically, three analyses were conducted to determine the extent to which fidelity growth was predicted by the number of years implementing, school contextual variables, and sustainability factors. Similar to research questions one and two, uninformative prior distributions were used for the 2008-09 sample and informative prior distributions were used for subsequent samples to demonstrate serial updating of prior distributions. Further, ML models were estimated to make comparisons and understand the likelihood's influence on posterior estimates. Also model fit diagnostics are presented to understand the extent to which predictors improved model generalizability. For the school contextual and years implementing analyses two level models were specified. For the years implementing analysis, the model was

$$\begin{aligned} &\text{Level 1:} \quad y_{tj} = \pi_{0j} + \pi_{1j}*time_t + e_{tj} \\ &\text{Level 2:} \quad \begin{aligned} \pi_{0j} &= \beta_{00} + \beta_{01}*x_1 + u_{0j} \\ \pi_{1j} &= \beta_{10} + \beta_{11}*x_1 \ + u_{1j}, \end{aligned} \end{aligned} \tag{7}$$

where x_1 is the number of years implementing. For the contextual analysis, the model with fixed effects predicting growth parameters was

$$\begin{aligned} &\text{Level 1:} \quad y_{tj} = \pi_{0j} + \pi_{1j}*time_t + e_{tj} \\ &\text{Level 2:} \quad \begin{aligned} \pi_{0j} &= \beta_{00} + \beta_{01}*x_1 + \beta_{02}*x_2 + \beta_{03}*x_3 \ + u_{0j} \\ \pi_{1j} &= \beta_{10} + \beta_{11}*x_1 + \beta_{12}*x_2 + \beta_{13}*x_3 \ + u_{1j}, \end{aligned} \end{aligned} \tag{8}$$

where x_1 is school size, x_2 is relative socio-economic status, and x_3 is a locality indicator. For the analysis of sustainability factors predicting fidelity growth parameters, a three level model was specified to take into account both school and district level sustainability predictors as follows:

$$\begin{aligned} &\text{Level 1:} \quad y_{tjk} = \pi_{0jk} + \pi_{1jk}*time_t + e_{tjk} \\ &\text{Level 2:} \quad \begin{aligned} \pi_{0kj} &= \beta_{00k} + \beta_{010}*x_1 + \beta_{020}*x_2 + u_{0j} \\ \pi_{1jk} &= \beta_{10k} + \beta_{110}*x_1 + \beta_{120}*x_2 + u_{1j} \end{aligned} \\ &\text{Level 3:} \quad \begin{aligned} \beta_{00k} &= \gamma_{000} + \gamma_{001}*x_3 + \gamma_{002}*x_4 + r_{0k} \\ \beta_{10k} &= \gamma_{100} + \gamma_{101}*x_3 + \gamma_{102}*x_4 + r_{1k}, \end{aligned} \end{aligned} \tag{9}$$

where x_1 and x_2 are school predictors for school priority and team use of data; and x_3 and x_4 are district predictors for district priority and capacity building. The model depicted in Equation 9 has the added complexity of variance for the intercept and slope terms (i.e., the fidelity growth parameters) at both the school and district level. For the school level, the Tao variance matrix is identical to Equation 6. For the district level, the Tao variance matrix is depicted in Equation 10:

$$T = \begin{bmatrix} \text{var}(r_{0k}) & - \\ \text{cov}(r_{0k}, r_{1k}) & \text{var}(r_{1k}) \end{bmatrix}. \tag{10}$$

CHAPTER IV

RESULTS

Descriptive Statistics

Descriptive statistics for outcome and predictive variables used in subsequent statistical models are provided in Table 3. The table provides descriptive information for each of the five samples and the total sample that includes all observations for all years. For 2010-11, descriptive statistics are provided for two samples. The sample with 10 schools is a subsample of the larger 2010-11 sample that was used for the demonstration of sustainability predictors influence on fidelity of implementation growth. This sample was notably smaller than the full sample as only a few schools met inclusion criteria for that analysis that required having three or more TIC scores and SUBSIST data for the 2010-11 school year.

TIC scores were used as the outcome variable in all analyses, and represented fidelity of program adherence. Sample means ranged from 21.79 to 28.65 over the five samples with the pooled sample having a mean of 23.83. Sample standard deviations also ranged from 5.21 to 8.27 with the pooled sample having a standard deviation of 7.32. The low sample standard deviation and high sample mean for the 2008-09 sample indicated that this sample was comprised of schools with overall higher fidelity of implementation scores and less variability. Consequently, it can be concluded that overall schools in this sample had a higher level of implementation fidelity than schools in other samples.

Descriptive statistics of TIC scores at the initial measurement occasion for each school are also provided. Overall, the figures reveal a high amount of variability at the

initial measurement occasion with means ranging from 15.97 to 26.38 across samples and standard deviations ranging from 7.29 to 9.71 across samples. Given this variability across samples and the procedure for coding time where zero represented the fidelity status of each school at the initial measurement occasion, substantive conclusions about fidelity at initial status and predictors of that status are limited, a topic further addressed in the discussion section.

Years implementing varied across the five samples. For most samples, the majority of schools were in their first or second year implementing. For the 2010-11 sample of schools used in the analysis of sustainability factors influence on fidelity growth, all ten schools were in year zero of SWPBIS implementation. This indicates that the schools in this sub-sample were putting critical pieces of SWPBIS in place, but had not begun implementing the program with students at the classroom level. Given the overall high mean of TIC scores for this 2010-11 subsample, these schools most likely had systems pieces in place at the end of the 2010-11 school year and were ready to begin implementing SWPBIS with students in 2011-12.

Means and standard deviations for sustainability factor scores estimated using the SUBSIST confirmatory factor analysis model (described in the measures section) are provided for the 2010-11 sub-sample and 2012-13 full sample of schools. The descriptive statistics indicated that school priority and district priority had similar distributions across the two samples. The standard deviations for school priority and district priority were a bit higher in 2010-11 probably due to the small sample size. The distributions for team use of data and capacity building differed between the two samples.

Table 3

Descriptive Statistics for Outcome Variable and Covariates by Each Sample Year, A Subset of the 2010-11 Sample with Sustainability Data, and for a Pooled Sample with Schools from All Years

	2008-09 (j = 13)	2009-10 (j = 61)	2010-11 (j = 53)	2010-11** (j = 10)	2011-12 (j = 64)	2012-13 (j = 85)	Total (j = 276)
Fidelity of implementation (TIC)							
Observations							
n for all schools	43	227	174	31	221	286	951
M per school	3.31	3.72	3.28	3.11	3.45	3.36	3.45
SD per school	0.48	0.95	0.66	0.33	0.53	0.57	0.69
All observations							
M	28.65	25.58	24.49	26.61	21.79	22.89	23.83
SD	5.21	6.66	6.83	6.04	8.27	6.94	7.32
Mdn	30.0	26.0	26.0	29.0	23.0	24.0	25.0
Min	10	4	3	11	0	3	0
Max	34	34	34	34	34	34	34
Initial observation							
M	26.38	21.80	18.72	23.80	15.97	18.38	19.02
SD	7.38	8.26	7.29	7.73	9.71	8.62	8.84
Mdn	28.0	23.0	19.0	22.5	13.5	19.0	19.0
Min	10	4	3	11	0	3	0
Max	34	34	34	34	34	34	34
Version							
2.2	43	202	4	1	0	0	249
3	0	25	169	29	0	0	194
3.1	0	0	1	1	221	286	508
Years implementing							
0	4	22	13	10	6	10	55
1	6	17	28	0	38	35	124
2	3	20	9	0	16	34	82
3	0	2	3	0	4	6	15
M	0.92	1.03	1.04	0.00	1.28	1.42	1.21
SD	0.76	0.91	0.81	0.00	0.72	0.79	0.82
Sustainability factor scores (SUBSIST)*							
School priority							
M				-0.17		-0.18	
SD				0.46		0.38	

Table 3 (continued)

	2008-09 (j = 13)	2009-10 (j = 61)	2010-11 (j = 53)	2010-11** (j = 10)	2011-12 (j = 64)	2012-13 (j = 85)	Total (j = 276)
Team use of data							
M				-0.35		-0.27	
SD				0.60		0.53	
District priority							
M				-0.16		-0.15	
SD				0.45		0.40	
Capacity building							
M				-0.25		-0.17	
SD				0.49		0.50	

Note. *For the 2010-11 school year, only 10 schools completed the SUBSIST measure. Subsequently, descriptive statistics are reported only for these 10 schools from 2010-11. **SUBSIST sample for the 2010-11 school year (a subset of the 2010-11 sample).

Research Question 1: Fidelity Growth Estimates for the 2008-09 Sample

Table 4 documents the results for the first research question addressing fidelity of implementation growth for the 2008-09 sample. Several different Bayesian models were considered all with different options for specifying *uninformative* prior distributions. I specified model numbers to facilitate comparison of the same model across samples. Bayesian model 1.0 utilized the uniform distributions for all model parameters, while other models used various combinations of conjugate distributions for priors. For example, model 2.0 used the normal distribution for fixed effects and the gamma distribution for all variance components. The hyper-parameters specifying the size of each prior distribution are also provided. It should also be noted that each model was estimated using MCMC sampling using the JAGS program (Plummer, 2003) using four Markov chains each allowed to iterate 20,000 times with the initial 10,000 iterations discarded from each chain and a thinning rate of 10. The $\hat{R}$ statistic for every model parameter in all models detailed in Table 4 was below 1.1 indicating adequate model convergence (Gelman & Hill, 2007).

Model fit criteria and reliability information were also provided to allow comparisons between models. Model reliability is defined in terms of the correlation of the predicted results to the observed results. Two reliabilities are provided for model fixed effects. The first is the correlation between the predicted and observed value. For the intercept, the correlation is between the predicted value for π_{0j} and the observed value at time zero for school zero. For the slope, the correlation is between each observed value after time zero and the predicted value calculated by multiplying π_{1j} by the time

variable. The final reliability statistic is a reliability coefficient calculated using a formula provided by Raudenbush and Bryk (2002, p. 49).

Based on model selection criteria, the overall best fitting Bayesian model is model 4.0 that utilized a normal distribution for the fixed effects, a gamma distribution for level one error variance, and a Wishart distribution for the level-two variance. Model code is provided in Appendix E. The AIC, BIC, and DIC for this model was lower than alternative models. Given that all of the models have the same number of parameters, a change in deviance test was not possible. Calculating the approximate Bayes factor using Equation 4 to compare model 4.0 to the models with the next lowest BICs, 3.0 and 2.0, resulted in values of 2.24 and 2.58 respectively. Using Jeffrey's recommendation that values over 1.0 support one model over another (Jeffreys (1961) as cited in Gill, 2002, p. 242), these results supported model 4.0 over 3.0 and 2.0. An additional consideration is model results that exceed acceptable ranges for results. For example, model 2.0 had a 95% credibility interval of the posterior distribution for the correlation between growth parameters of (.06, 1.04). The high end of the credibility interval is not possible given that correlations cannot exceed an absolute value of 1.0. Considering all of these criteria converge on supporting model 4.0 over alternative models, results for this model will be interpreted.

Model results for Bayes model 4.0 are as follows. The fixed effect for the intercept, β_{00}, represented an average fidelity of implementation value at time zero and had a posterior distribution 95% credibility interval of (22.49, 30.66). The fixed effect for the slope, β_{10}, represented the average change in fidelity scores for each month implementing SWPBIS during that school year and had a posterior distribution 95%

credibility interval of (-0.03, 1.20). Considering that the credibility interval for posterior distribution of the fixed effect of the slope included zero indicating that this model parameter may not have been an influential predictor for this sample. Given the relatively high fidelity scores as explicated in the descriptive statistics section, a ceiling effect may be occurring as schools in the 2008-09 sample on average had high TIC score and did not have much room to gain. Using the posterior distribution mean for the random effects to calculate intra-class correlations (ICCs) revealed that 96.8% of the variance was between schools while the remaining 3.2% was within schools between measurement occasions. The correlation estimate between the intercept and slope parameters revealed that the growth parameters were negatively correlated with a 95% credibility interval for the posterior distribution ranging from (-.95, -.48). This implies that across all schools, as the intercept term gets larger, the slope gets smaller and as the intercept gets smaller, the slope gets larger. Intuitively, this is logical as schools with a smaller initial fidelity score have more room to grow.

Model results for Bayes model 4.0 were similar to the ML model for the 2008-09 sample data. This is logical as the Bayes model accounts for the likelihood distribution and prior distribution, and the prior distribution was not specific given the use of uninformative conjugate prior distributions for model parameters. The 95% *confidence* interval for ML model estimates is analogous to the 95% *credibility* interval for Bayesian posterior distribution estimates. The *confidence* interval of the slope of the ML model is the parameter estimate +/- 1.96 times the standard error, or (0.05, 1.16). As can be seen, this interval is very similar to the 95% credibility interval for the slope of Bayes model

4.0. Slight deviations between the Bayes model and the ML model are on account of variability associated with MCMC sampling.

Research Question 2: The Influence of Bayesian Updating on Fidelity Growth Estimates

As with the previous research question, many model possibilities were explored. Bayesian models were specified using both uninformative and informative prior distributions with varying shapes in sizes to explore the impact of model specification on generalization criteria including AIC, BIC and DIC. In terms of varying the shape of prior distributions, uninformative models were estimated for the 2009-10 sample using specifications for models 1.0 and 4.0 detailed in the previous section. Information criteria for model 4.0 (AIC = 1019.04, BIC = 1031.70, DIC = 1278.35) were lower than model 1.0 (AIC = 1052.41, BIC = 1065.08, DIC = 1555.73) for all values. Further, calculating the approximate Bayes factor between model 4.0 and 1.0 for the 2009-10 sample resulted in a value well above 1.0 providing support for model 4.0 over 1.0. Subsequently, model 4.0 was used as a base model for the subsequent analyses.

Building on model 4.0, variations of hyper-parameters for model parameter prior distributions were explored for the four samples collected between 2009-10 and 2012-13. Information criteria for the various Bayesian and ML models are highlighted in Table 5. For all models detailed, the random effects were specified using uninformative priors. Specifically, a gamma distribution with hyper-parameters (1, 1) was used for the level-one error variance, and the Wishart distribution with hyper-parameters $df = 3$ was used for the level-two growth parameter variance. For the fixed effects a normal distribution was used with the following hyper-parameters: (a) *Uninformative* prior distributions use

Table 4

Fidelity Growth Estimates Based on Various Estimates for a 2008-09 Sample of Schools (J = 13) Implementing School-Wide Positive Behavior Interventions and Supports

	β_{00}	β_{10}	$\sigma_{\pi_{0j}}$	$\sigma_{\pi_{1j}}$	$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	$\sigma_{e_{tj}}$	AIC	BIC	DIC	Model Reliability
Bayes model 1.0							154.651	158.040	256.528	0.987
Prior form	uniform	uniform	uniform	uniform	uniform	uniform				
Hyper-parameters	(0, 34)	(-34, 34)	(0, 1000)	(0, 1000)	(-1, 1)	(0, 1000)				
M	26.470	0.643	7.728	1.105	-.012	1.411				
SD	2.264	0.328	1.943	0.297	.573	0.315				
Med	26.488	0.645	7.376	1.065	-.015	1.363				
95% CI*	(21.7, 31.0)	(0.0, 1.3)	(5.0, 12.5)	(0.7, 1.8)	(<-.9, .9)	(1.0, 2.1)				
Reliability**	0.998	0.975								
Reliability***		0.811								
Bayes model 2.0							152.697	156.086	245.624	0.987
Prior form	normal	normal	gamma	gamma	gamma	gamma				
Hyper-parameters	(0, 10000)	(0, 10000)	(1, 1)	(1, 1)	(1, 1)	(1, 1)				
M	26.419	0.632	6.700	1.040	.279	1.349				
SD	2.012	0.330	1.448	0.248	.353	0.323				
Med	26.415	0.630	6.486	1.009	.185	1.300				
95% CI*	(22.6, 30.5)	(0.0, 1.2)	(4.5, 10.1)	(0.7, 1.6)	(.1, >.9)	(1.0, 2.0)				
Reliability**	0.998	0.975								
Reliability***		0.794								
Bayes model 3.0							152.413	155.802	248.069	0.987
Prior form	normal	normal	gamma	gamma	uniform	gamma				
Hyper-parameters	(0, 10000)	(0, 10000)	(1, 1)	(1, 1)	(-1, 1)	(1, 1)				
M	26.479	0.639	6.696	1.035	-.016	1.344				
SD	2.007	0.328	1.470	0.249	.584	0.341				
Med	26.516	0.638	6.487	0.997	-.021	1.299				
95% CI*	(22.6, 30.3)	(0.03, 1.26)	(4.5, 10.1)	(0.7, 1.6)	(<-.9, >.9)	(1.0, 2.0)				
Reliability**	0.998	0.975								
Reliability***		0.791								

Table 4 (continued)

	β_{00}	β_{10}	$\sigma_{\pi_{0j}}$	$\sigma_{\pi_{1j}}$	$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	$\sigma_{e_{tj}}$	AIC	BIC	DIC	Model Reliability
Bayes model 4.0							150.799	154.188	220.270	0.987
Prior form	normal	normal	Wishart	Wishart	Wishart	gamma				
Hyper-parameters	(0, 10000)	(0, 10000)	df = 3	df = 3	df = 3	(1, 1)				
M	26.556	0.602	7.177	1.031	-.794	1.313				
SD	2.055	0.307	1.576	0.241	.121	0.242				
Med	26.541	0.599	6.930	0.996	-.821	1.275				
95% CI*	(22.5, 30.7)	(0.0, 1.2)	(4.9, 10.9)	(0.7, 1.6)	(<-.9, -.5)	(0.9, 1.9)				
Reliability**	0.998	0.975								
Reliability***		0.793								
Bayes model 5.0							154.633	158.022	245.982	0.987
Prior form	normal	normal	Wishart	Wishart	Wishart	uniform				
Hyper-parameters	(0, 10000)	(0, 10000)	df = 3	df = 3	df = 3	(0, 1000)				
M	26.521	0.599	7.114	1.014	-.811	1.405				
SD	2.080	0.307	1.537	0.237	.116	0.291				
Med	26.534	0.605	6.871	0.990	-.838	1.360				
95% CI*	(22.5, 30.7)	(0.0, 1.2)	(4.8, 10.8)	(0.6, 1.6)	(<-.9, -.5)	(1.0, 2.1)				
Reliability**	0.998	0.974								
Reliability***		0.777								
ML model							229.376	239.943		0.987
Estimate	26.511	0.604	6.893	0.978	-.874	1.269				
SE	1.939	0.284								
t-value	13.676	2.128								
Reliability**	0.998	0.974								
Reliability***		0.776								

Notes. Model reliability is correlation between observed scores and predicted scores. *For Bayesian estimates, CI refers to a credibility interval based on Gibbs sampling. **Reliability calculated as (a) correlation between observed value at time 0 and intercept estimate for each school, and (b) correlation between observed change in TIC score and predicted change. ***Reliability calculated according to equations from Raudenbush and Bryk (2002).

the same values specified for model 4.0 used in the 2008-09 sample, (b) s*omewhat* informative priors were specified using the mean of the previous sample's posterior distribution and three times the posterior standard deviation, (c) *informative* priors were specified as the mean of the previous samples' posterior distribution and two times the standard deviation, and (d) *very informative* were specified as the mean and standard deviation of the previous sample's posterior distribution. Like the previous research question, each model was estimated using MCMC sampling using the JAGS program (Plummer, 2003) using four Markov chains each allowed to iterate 20,000 times with the initial 10,000 iterations discarded from each chain and a thinning rate of 10. Unless otherwise noted, the $\hat{R}$ statistic for every model parameter in all models detailed in Tables 5 and 6 were below 1.1 indicating adequate model convergence (Gelman & Hill, 2007).

For the 2009-10 analyses, evidence in Table 5 reveals that information criteria were lowest for the *very informative* model. The AIC, BIC and DIC were lowest for this model, but the AIC and BIC are very close to the next lowest model specified with *informative* priors. As was the case for the first research question, change in deviance tests cannot be performed as these models have the same number of parameters. The approximate Bayes factor comparing the *very informative* model to the *informative* model was 1.05 giving support for the *very informative* model. If I were to only report on this sample, I would report the results of the *very informative* model given the convergence of model selection criteria described. I would, however, compare model results to the *uninformative* model to make a qualitative judgment about the extent to which the prior distribution influenced posterior distributions for parameters of interest. In the end, I am not interested the results of the 2009-10 sample in isolation, but did use the results for the

very informative model as the basis for prior distribution hyper-parameters for the 2009-10 sample.

In this regard, prior distributions for the fixed effects were updated for each informative model in a serial fashion. Specifically, the model that had the overall best fit for each sample was used to form the informative prior distributions for subsequent year's analyses. Considering the 2009-10 results discussed above, the posterior distributions for the *very informative* model were used as the prior distributions of the fixed effects for the analyses involving the 2010-11 sample. As with 2009-10 sample, subsequent sample analyses used a varying range of hyper-parameters from *somewhat informative* to *informative* to *very informative*.

For the 2010-11 analysis, evidence in Table 5 model selection evidence converged on the *somewhat informative* model. The AIC, BIC and DIC for this model is the lowest as compared to other models. Further, the approximate Bayes factor when comparing this model to the model with the next lowest BIC, the *uninformative* model, was 2.01. Given this evidence, the posterior distributions for the *somewhat informative* model were used as prior distributions for the 2011-12 analysis.

For the 2011-12 analysis, model selection evidences converged on the *somewhat informative* model. The AIC and BIC was lowest for this model. The DIC, however, was lower for the *uninformative* and *informative* models. The approximate Bayes factor comparing the *somewhat informative* to the model with the next lowest BIC, the *very informative* model, was 1.65 providing a third piece of evidence in support of the *somewhat informative* model. It should be noted that the all four Bayesian models for the 2011-12 sample did not converge based on the criterion of $\hat{R}$ less than 1.1 for all model.

Table 5
Information Criteria for Multilevel Models of Fidelity Growth

	AIC	BIC	DIC
2009-10			
ML model	1261.140	1281.689	
Bayesian models			
Uniformative	1019.036	1031.702	1278.352
Somewhat informative	1018.430	1031.095	1280.575
Informative	1018.367	1031.033	1275.954
Very informative	1018.271	1030.936	1263.778
2010-11			
ML model	1061.359	1080.313	
Bayesian models			
Uniformative	959.426	971.248	1148.105
Somewhat informative	958.032	969.854	1133.777
Informative*	964.276	976.098	1146.783
Very informative	959.813	971.635	1156.943
2011-12			
ML model	1363.315	1383.704	
Bayesian models			
Uniformative*	1187.611	1200.564	1349.384
Somewhat informative*	1181.781	1194.735	1353.810
Informative*	1183.028	1195.982	1352.070
Very informative*	1182.785	1195.739	1361.759
2012-13			
ML model	1741.231	1763.167	
Bayesian models			
Uniformative	1547.398	1562.054	1824.731
Somewhat informative	1540.862	1555.518	1821.736
Informative	1545.938	1560.594	1839.226
Very informative*	1541.560	1556.216	1817.725

Note. The shape of the prior distributions was normal for the fixed effects, a gamma distribution for level 1, and Wishart for level 2. *These models did not converge using the criterion of $\hat{R} < 1.1$ for all parameters.

parameters. For all of the models in this year's sample, $\hat{R}$ was greater than 1.1 for the correlation between growth parameters, $\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$. Since this parameter was not used to form the prior distribution for a subsequent model, and all other model parameters had a $\hat{R}$ less than 1.1, this non-convergence is less of a concern, but limitations are discussed. Given that three out of four criteria supported the *somewhat informative* model, this

model's posterior distributions for the fixed effects were used as prior distributions in the 2012-13 analysis

For the 2012-13 analysis, model selection evidenced converged on the *somewhat* informative model. The AIC, BIC, and DIC were lowest for this model. Further, the approximate Bayes factor comparing the *somewhat informative* model to the model with the next lowest BIC, the *very informative* model, was 1.42 providing evidence for the *somewhat informative* model. The results for the *somewhat informative* 2012-13 model are highlighted below.

Table 6 highlights the results for the selected models for all years that were incorporated as prior distributions in subsequent years. Specifically, results are reported for the *very informative* model of the 2009-10 sample and the *somewhat informative* models for the 2010-11, 2011-12, and 2012-13 samples. The 2012-13 model is of particular interest as model-based inferences are contingent on the data for that year as well as the results from all the previous samples. Subsequently, the results are the most-up-to date parameter estimates representing the most current version of the *true* parameters of interest (Gelman & Shalizi, 2013), the fidelity growth parameters for SWPBIS.

As the 2012-13 represented the most-up-to-date knowledge of fidelity growth, this model is interpreted. The average fidelity score for schools implementing SWPBIS at the first measurement occasion had a posterior distribution mean of 20.65 with a 95% credibility interval for the posterior distribution of (19.64, 21.69). The average fidelity growth per month of implementation after the initial measurement occasion had a posterior distribution mean 0.83 with a 95% credibility interval for the posterior

Table 6
Bayesian Multilevel Fidelity of Implementation Growth Models

	2009-10 (j = 61)	2010-11 (j = 53)	2011-12 (j = 64)	2012-13 (j = 85)
Fixed Effects				
Intercept: β_{00}	25.22 (0.60)	23.29 (0.64)	21.76 (0.66)	20.65 (0.53)
Time (slope): β_{10}	0.68 (0.09)	0.95 (0.11)	0.80 (0.09)	0.83 (0.08)
Random Effects				
$\sigma_{\pi_{0j}}$	8.02 (0.81)	6.24 (0.89)	9.31 (0.95)	6.92 (0.64)
$\sigma_{\pi_{1j}}$	0.81 (0.11)	0.61 (0.16)	0.69 (0.11)	0.73 (0.10)
$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	-.86 (.05)	-.84 (.10)	-.92 (.05)	-.91 (.04)
$\sigma_{e_{tj}}$	2.25 (0.15)	3.72 (0.29)	3.45 (0.22)	3.53 (0.21)
Intra-class correlations				
Level 1	.072	.261	.120	.205
Level 2	.928	.739	.880	.795

Note. For model values, the number outside the parentheses refers to the mean of the posterior distribution, and inside the parentheses is the posterior standard deviation. *j* refers to the number of schools in that year's sample.

distribution of (0.67, 0.99). This indicated that over a three-month period, schools gained about two to three points on the TIC, which equates to implementing two to three critical components of SWPBIS. ICCs of the variance components showed that the majority, 79.5%, of the variance was between schools. The correlation of the growth parameters was negative, -.91, with a 95% credibility interval of (-.97, -.81). This negative correlation implies that as the intercept term increased, the slope decreased as schools approached the maximum score on the TIC. Conversely, schools with lower initial fidelity scores had higher growth.

The Bayesian models were also compared to the ML models and ML model results are presented in Table 7. It is important to look at the ML results for two main reasons. First, the likelihood distribution is incorporated into the Bayesian estimate.

Second, the ML estimates provides insight into relative influence of the observed data and prior distribution into Bayesian posterior distributions.

Table 7
Maximum Likelihood Multilevel Fidelity of Implementation Growth Models

	2009-10 (j = 61)	2010-11 (j = 53)	2011-12 (j = 64)	2012-13 (j = 85)	Pooled (j = 276)
Fixed Effects					
Intercept: β_{00}	22.24 (0.99)	20.10 (0.88)	17.40 (1.1)	19.19 (0.81)	20.00 (0.48)
Time (slope): β_{10}	0.95 (0.11)	1.26 (0.12)	1.12 (0.1)	0.98 (0.10)	1.04 (0.05)
Random Effects					
$\sigma_{\pi_{0j}}$	7.47	5.54	8.31	6.83	7.46
$\sigma_{\pi_{1j}}$	0.75	0.53	0.59	0.71	0.67
$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	-.87	-.87	-.96	-.93	-.89
$\sigma_{e_{tj}}$	2.24	3.71	3.50	3.54	3.24
Intra-class correlations					
Level 1	.082	.308	.150	.210	.158
Level 2	.918	.692	.850	.790	.842

Note. For model values, the number outside the parentheses refers to the parameter estimate, and inside the parentheses is the standard error. j refers to the number of schools in that year's sample.

For the fixed effect of the intercept, β_{00}, the Bayesian estimates with informative priors and the ML estimates differ. The difference is on account of the influence of the prior distribution on posterior estimates. Figure 3 depicts the mean intercept estimate for the Bayesian models and ML estimate for the slope over the five samples. For the 2008-09 sample the intercept estimates are about equal because of the uninformative prior distribution used for the Bayesian model. For the other samples, the Bayesian estimates tended to be higher than the ML estimates. This pattern is occurring because of the prior distributions' influence on posterior estimates. Referring to descriptive statistics of the initial observation reported in Table 3, it can be seen that the ML estimates more closely resemble the observed score mean at time zero as can be expected given the weight the prior distribution has on the model results. As can be observed in the image, the range in

estimates for the intercept across samples is smaller for the Bayesian models than the ML models.

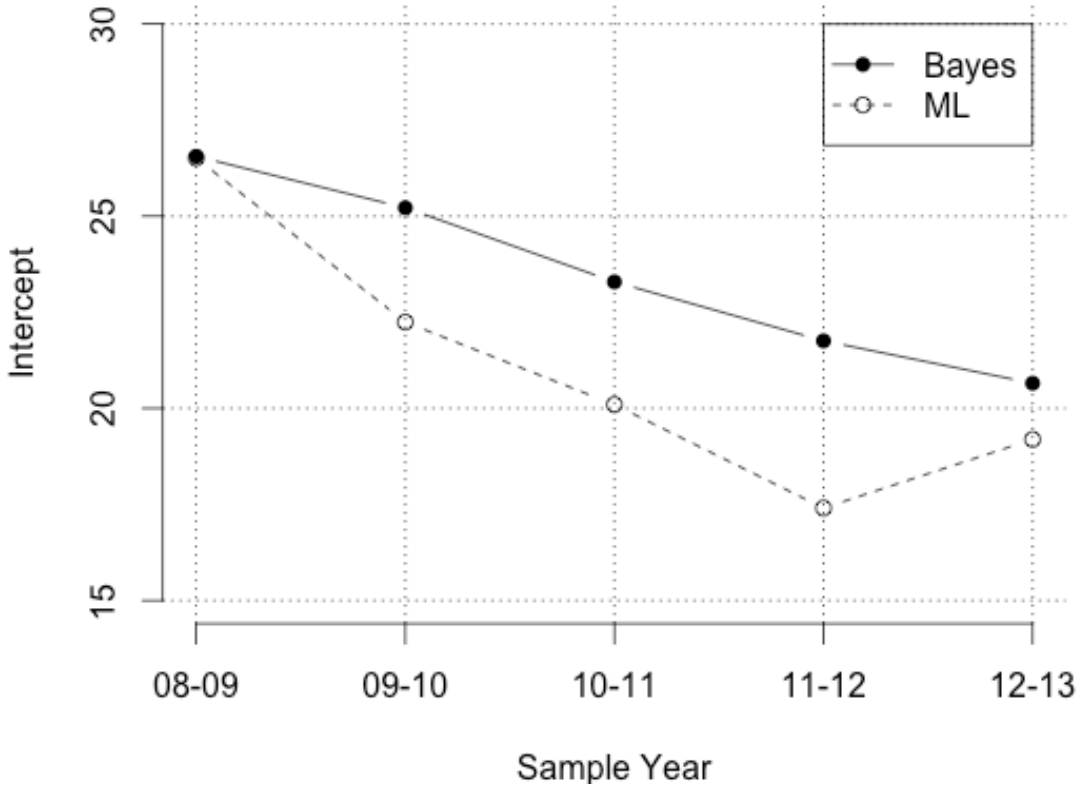

Figure 3. Intercept estimates based on the mean of the Bayesian posterior distribution and maximum likelihood (ML) estimates.

For the fixed effect of the slope, β_{10}, the Bayesian and ML estimates differ as well. Figure 4 depicts the average slope estimates for each year. Conversely to the intercept, the average slope estimates for the Bayesian models were lower than the ML model. Like the intercept, the differences in slope estimates are on account of prior distributions' influence on posterior distributions. Like the intercept, the range in estimates across samples for the mean of the Bayesian models is smaller than the ML models.

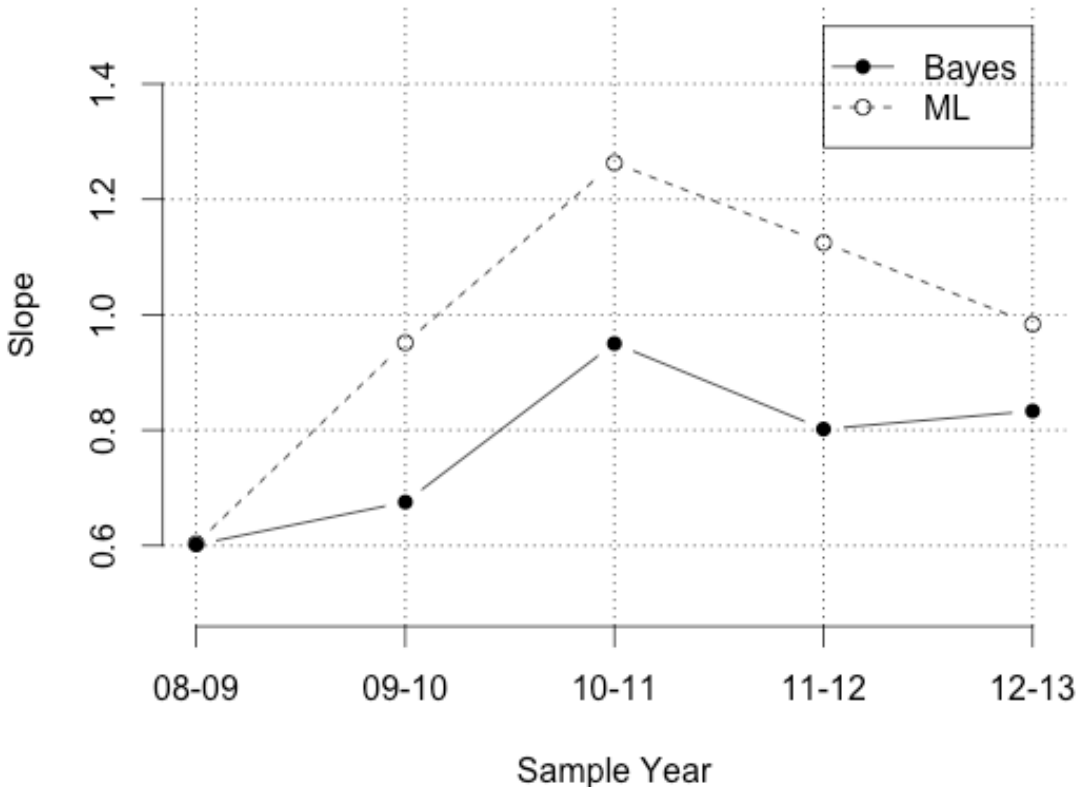

Figure 4. Slope estimates based on the mean of the Bayesian posterior distribution and maximum likelihood (ML) estimates.

Results from a pooled model estimated with ML are also displayed in Table 7. The pooled model takes into account all observations collected over all years. The average intercept estimate for this model was 20.00, and the average intercept was 1.04. While numerical comparisons between the pooled ML model and 2012-13 Bayesian model cannot be made, qualitative differences between models can be described. For example, the intercept estimate for the pooled ML model was within the range of the 95% credibility interval for the posterior distribution of the 2012-13 Bayesian model intercept. The slope estimate for the pooled ML model was not within the 95% credibility interval for the posterior distribution of the 2012-13 Bayesian model.

Model reliability indexes were also calculated and are displayed in Table 8. Several correlations were calculated to compare the observed data to predicted results. The model correlation was calculated by taking the correlation between the predicted TIC score based on model parameters and the observed score. Overall, the model correlations for the Bayesian models were slightly (perhaps negligibly) higher than the ML models, and in the relatively high range (.909 to .987). The intercept correlation was calculated by taken the correlation between the observed score at time zero for each school and the school's estimated intercept. Like the model correlation, the intercept correlation was slightly higher for the Bayesian models, and within a relatively high range across samples (.960 to .998). The slope correlation was calculated by taking the correlation between the observed change in TIC scores and the predicted change. Predicted change was calculated by multiplying each school's slope parameter, π_{1j}, by the time variable. Again, the Bayesian models' slope correlations were higher than the ML models' with a range between .676 and .975.

Table 8
Reliability Indexes for Bayesian and Maximum Likelihood Models Across Samples

	2008-09 (j = 13)	2009-10 (j = 61)	2010-11 (j = 53)	2011-12 (j = 64)	2012-13 (j = 85)
Bayesian model					
Model correlation	.987	.967	.909	.944	.920
Intercept correlation	.998	.975	.960	.976	.963
Slope correlation	.975	.858	.676	.755	.807
Slope reliability	.793	.544	.233	.298	.309
ML model					
Model correlation	.987	.966	.898	.937	.915
Intercept correlation	.998	.974	.955	.972	.962
Slope correlation	.974	.852	.616	.707	.790
Slope reliability	.776	.516	.193	.240	.295

Note. Intercept correlation calculated as the correlation between observed value at time 0 and intercept estimate for each school. Slope correlation calculated as the correlation between observed change in TIC score and predicted change. Slope reliability calculated according to hierarchical model parameter reliability formula provided by Raudenbush and Bryk (2002).

Slope reliability was calculated using the following formula provided by Raudenbush and Bryk (2002, p. 50):

$$reliability(\hat{\beta}_{10}) = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{\tau_{11}}{\tau_{11} + v_{11j}}\right), \quad (11)$$

where $\hat{\beta}_{10}$ is the average slope parameter estimate; τ_{11} is the parameter variance for the slope as specified in variance matrix of the model outlined in Equation 6; and v_{11j} is the residual variance for the slope estimate for school *j*. The residual variance was calculated by (a) calculating the observed difference between the observed score at $time_t$ and the initial score at $time_t = 0$; (b) calculating the predicted difference at $time_t$ by multiplying school *j*'s slope parameter, π_{1j}, by $time_t$; (c) calculating a residual term for each observation by subtracting the predicted difference from the observed difference; (d) calculating the residual variance for school *j* by taking the variance of the residual for all observations from school *j*; and (e) dividing the variance by the number of observations for school *j*. Slope reliabilities were slightly higher for the Bayesian than ML models of each sample. Besides the model for the 2008-09 sample, the slope reliabilities of the Bayesian models were low, ranging from .23 to .54 for the four samples collected between 2009-13. These reliabilities were low due to the way the formula specifies calculating the residual variance for the slope, v_{11j}. Specifically, schools typically had three to four TIC scores, resulting in a higher v_{11j} estimate than if they had many more observations. If more observations had occurred per year, the denominator of Equation 11 would be expected to be lower, resulting in a higher reliability index.

Research Question 3: Demonstrating Adding Fixed-Effects Predictors for Bayesian Models Using Bayesian Updating Methodology

In the following section, three separate analyses will demonstrate how predictors can be included in analyses, and how to apply serial updating to the fixed effects. The first example will explore the extent to which the number of years implementing SWPBIS predicts SWPBIS fidelity growth. The second example will explore the extent to which contextual variables including school size, relative socio-economic status, and locality predict SWPBIS fidelity growth. Finally, the last example will explore sustainability factors' predictive influence on fidelity growth.

Years implementing predicting fidelity growth? Years implementing offers an interesting opportunity to demonstrate the inclusion of one school level variable as predictors of growth parameters. The model used for this demonstration is highlighted in Equation 7 and was estimated using model code available in Appendix F. Descriptive information about the outcome variable, TIC fidelity scores, and the predictor variable, years implementing, is available in Table 3. The models for this analysis were estimated using JAGS (Plummer, 2003) for the R program (R Core Team, 2012) and four Markov chains . For these models, the number of iterations was 100,000 with a thinning rate of 80 to facilitate model convergence. Unless otherwise indicated, the models reported below had $\hat{R}$ values below 1.1 for all parameters indicating adequate model convergence. All models were specified using normal distribution for fixed effect prior distributions, gamma distributions for level-1 error variance, and the Wishart distribution for level-2 variance.

Using a fully Bayesian approach incorporating informative prior distributions for fixed effects, previous sample results were used as prior distributions in each subsequent sample. To this end, the 2008-09 sample was estimated using uninformative prior distributions, and subsequently the 2009-10 sample was estimated using the results from the 2008-09 results. The 2010-11 sample was estimated using the best fitting results from the 2009-10 analyses that including fitting models specified with uninformative prior distributions, *very informative* priors, *informative* priors, and *somewhat informative* priors. *Very informative* to *somewhat informative* prior distributions were defined by the study author, and represented a range in specificity for the prior distribution. For each of the three categories, (a) the means of the fixed effects for the previous sample's results was used as the prior distribution mean; and (b) the standard deviation ranged from the exact standard deviation for *very specific* prior distributions, two times the standard deviation for *specific* priors, and three times for *somewhat specific* priors. All prior distributions for fixed effects were specified as normally distributed. In a serial fashion, the 2011-12 informative models were estimated using the best fitting results for 2010-11, and the 2012-13 informative models were estimated using the best fitting results from 2011-12.

Information criteria for covariate models with years implementing predicting growth parameters are highlighted in Table 9. Results from the 2008-09 uninformative model were used as the basis for fixed effect prior distribution hyper-parameters for the 2009-10 analyses. Results from the 2009-10 *very informative* model were used as the basis for prior distribution for the 2010-11 analyses based on three of four model fit criteria favoring this model over other models. For the 2009-10 analyses, the *very*

informative model had the lowest AIC and BIC, and the approximate Bayes factor comparing the *very informative* model to the model with the next lowest BIC, the *informative* model, was 1.02 favoring this model over other models. Results from the 2010-11 *somewhat informative* model were used as prior distribution hyper-parameters for the 2011-12 analysis as all four model selection criteria favored this model. The AIC, BIC, and DIC were lowest for this model, and the approximate Bayes factor comparing the *somewhat informative* model to the *very informative* model was 1.68. Result from the 2011-12 *somewhat informative* model were used as the hyper-parameters for the 2012-13 analysis because three of four model selection criteria converged on this model. The AIC and BIC were lowest for the *somewhat informative model* in 2011-12, and the approximate Bayes factor comparing the *somewhat informative* model to the *uninformative* model was 4.42. It should also be noted that the *informative* and *very informative* models did not adequately converge because not all model parameters had values for $\hat{R}$ less than 1.1. For the 2011-12 *informative* model, $\hat{R}$ for $\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$ was greater than 1.1 ($\hat{R}$ = 1.33), and for the *very informative* model most model parameters were above the 1.1 threshold with $\hat{R}$ values over 10 for several model parameters. Model selection criteria from the 2012-13 analyses suggested that the *informative* model had the best overall fit. The AIC, BIC, and DIC were lowest for this model, and the approximate Bayes factor comparing the *informative* model to the *somewhat informative* model was 1.26. Results from the 2012-13 *informative* model are interpreted below as they represented the most current understanding of years implementing influence on SWPBIS fidelity growth.

Table 9
Information Criteria for Multilevel Models with Years Implementing as Predictor

	AIC	BIC	DIC
2008-09			
ML model	232.89	246.98	
Bayesian models			
Uninformative	151.64	156.16	211.09
2009-10			
ML model	1236.22	1263.62	
Bayesian models			
Uninformative	1018.54	1035.43	1272.63
Somewhat informative	1017.04	1033.93	1266.69
Informative	1016.98	1033.87	1260.97
Very informative	1016.93	1033.82	1263.88
2010-11			
ML model	1063.31	1088.58	
Bayesian models			
Uninformative	961.03	976.79	1172.08
Somewhat informative	958.47	974.23	1145.08
Informative*	959.97	975.73	1148.10
Very informative	959.50	975.26	1151.44
2011-12			
ML model	1326.25	1353.44	
Bayesian models			
Uninformative	1186.00	1203.27	1369.69
Somewhat informative	1183.03	1200.30	1366.87
Informative*	1186.97	1204.24	1364.54
Very informative*	1331.23	1348.50	1502.87
2012-13			
ML model	1701.58	1730.83	
Bayesian models			
Uninformative	1548.19	1567.73	1835.34
Somewhat informative	1540.97	1560.51	1819.55
Informative	1540.51	1560.06	1812.96
Very informative	1547.20	1566.74	1845.12

Note. *These models did not converge using the criterion of $\hat{R} < 1.1$ for all parameters.

Results form the 2012-13 analysis as well as results from previous years that were used as prior distributions are presented in Table 10. The intercept and slope parameters, β_{00} and β_{10}, represented the average initial and change in fidelity estimates for schools in year zero of implementation. The average initial fidelity score for schools in year zero of implementation was 21.59 with a 95% credibility interval for the posterior distribution of (20.07, 3.17). The average change in fidelity score per month of school for schools in

year zero of SWPBIS implementation was 0.78 with a 95% credibility interval for the posterior distribution of (0.40, 1.16). Results indicated that years implementing did not predict a change in the initial status and or slope estimate. The posterior distribution for years implementing predicting initial status, β_{01}, had a mean of -0.34 and a 95% credibility interval of (-1.46, 0.72). The posterior distribution for years implementing predicting the slope, β_{11}, had a mean of 0.00 and a 95% credibility interval of (-0.22, 0.24). Since the credibility interval of both these parameters crossed zero, it can be concluded that years implementing did not predict initial or change in fidelity scores.

Table 10
Bayesian Multilevel Fidelity of Implementation Growth Models With Years Implementing as a Predictor and Updated Prior Distributions

	2008-09 (j = 13)	2009-10 (j = 61)	2010-11 (j = 53)	2011-12 (j = 64)	2012-13 (j = 85)
Fixed Effects					
Intercept: β_{00}	27.58 (3.5)	26.77 (0.5)	24.80 (0.7)	23.65 (0.7)	21.59 (0.8)
Yrs imp: β_{01}	-1.15 (3.0)	-0.85 (0.5)	-1.73 (0.6)	-1.72 (0.6)	-0.34 (0.6)
Time (slope): β_{10}	0.58 (0.5)	0.38 (0.2)	0.69 (0.2)	0.56 (0.2)	0.78 (0.2)
Yrs imp*time: β_{11}	0.05 (0.5)	0.20 (0.1)	0.28 (0.1)	0.20 (0.2)	0.00 (0.1)
Random Effects					
$\sigma_{\pi 0j}$	7.74 (1.8)	8.78 (0.9)	6.50 (0.9)	10.07 (1.0)	7.26 (0.7)
$\sigma_{\pi 1j}$	1.12 (0.3)	0.94 (0.1)	0.65 (0.2)	0.79 (0.1)	0.75 (0.1)
$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	-.74 (.2)	-.89 (.0)	-.83 (.1)	-.92 (.1)	-.91 (.0)
$\sigma_{e_{tj}}$	1.3 (0.2)	2.23 (0.2)	3.71 (0.3)	3.45 (0.2)	3.52 (0.2)
Intra-class correlations					
Level 1	.027	.058	.244	.104	.189
Level 2	.973	.942	.756	.896	.811

Note. For model values, the number outside the parentheses refers to the mean of the posterior distribution, and inside the parentheses is the posterior standard deviation. j refers to the number of schools in that year's sample.

Because Bayesian analyses incorporate the likelihood as part of the estimation algorithm, it is also important to examine these estimates. Subsequently, ML model estimates are proved in Table 11. Looking closely at the fixed effect estimates, differences can be seen comparing the ML estimates to the Bayesian estimates for samples collected between 2009 and 2013. Further, the ML estimates for years implementing are significant based on parameter *t*-tests. Given the difference in magnitude for parameter estimates and differences in patterns of influence, model based inferences based ML estimates would be quite different than those of the Bayesian models. The Bayesian model estimates appear to be largely influenced by the inclusion of the prior distribution. Of note, the Bayesian models with informative priors were formally compared using model selection criteria to both ML estimates and models with uninformative prior distributions, and in all cases the Bayesian models with informative priors were selected. This finding will be discussed further.

Does school context predict SWPBIS fidelity growth? Using contextual variables provided by a federal database about schools (National Center for Education Statistics, 2011) provides a demonstration of the influence of school context variables on fidelity growth parameters. The model used for this demonstration is highlighted in Equation 8 and was estimated using the same code as the demonstration with years implementing predicting growth (available in Appendix F). Descriptive informative for fidelity outcome and contextual variables is provided in Tables 1 and 3. Like the previous demonstration, models for this analysis were estimated using JAGS (Plummer, 2003) for the R program (R Core Team, 2012) and four Markov chains. For these

Table 11
Maximum Likelihood Multilevel Fidelity of Implementation Growth Models With Years Implementing as a Predictor

	2008-09 (j = 13)	2009-10 (j = 61)	2010-11 (j = 53)	2011-12 (j = 64)	2012-13 (j = 85)	Pooled (j = 276)
Fixed Effects						
Intercept: β_{00}	27.53 (3.1)***	16.97 (1.2)***	19.01 (1.4)***	6.64 (1.6)***	10.66 (1.3)***	14.76 (0.8)***
Yrs imp: β_{01}	-1.10 (2.6)	5.10 (0.9)***	1.05 (1.1)	8.41 (1.1)***	6.01 (0.8)***	4.34 (0.5)***
Time (slope): β_{10}	0.58 (0.5)	1.37 (0.2)***	1.27 (0.2)***	1.89 (0.2)***	1.92 (0.2)***	1.53 (0.1)***
Yrs imp*time: β_{11}	0.03 (0.4)	-0.40 (0.1)**	0.00 (0.2)	-0.60 (0.1)***	-0.65 (0.1)***	-0.41 (0.1)***
Random Effects						
$\sigma_{\pi 0j}$	6.85	5.88	5.47	5.73	4.88	6.55
$\sigma_{\pi 1j}$	0.98	0.66	0.53	0.41	0.48	0.58
$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	-.88	-.83	-.87	-.91	-.87	-.86
$\sigma_{e_{tj}}$	1.27	2.24	3.70	3.49	3.56	3.24
Intra-class correlations						
Level 1	.033	.125	.313	.270	.345	.195
Level 2	.967	.875	.687	.730	.655	.805

Note. For model values, the number outside the parentheses refers to the maximum likelihood parameter estimate, and inside the parentheses is the standard error. j refers to the number of schools in that year's sample. **$p < .01$. ***$p < .001$.

models, the number of iterations was 100,000 with a thinning rate of 80 to ensure adequate model convergence. Unless otherwise indicated, the models reported below had $\hat{R}$ values below 1.1 for all parameters indicating adequate model convergence. All models were specified using normal distribution for fixed effect prior distributions, gamma distributions for level-1 error variance, and the Wishart distribution for level-2 variance.

Like the previous demonstration, the analyses of contextual variables influence on fidelity growth incorporated informative prior distributions for the fixed effects to demonstrate a Bayesian updating process. The 2008-09 sample was estimated using

uninformative prior distributions, and subsequently the 2009-10 sample was estimated using the results from the 2008-09 results. The 2010-11 sample was estimated using the best fitting results from the 2009-10 analyses that including fitting models specified with uninformative prior distributions, *very informative* priors, *informative* priors, and *somewhat informative* priors. *Very informative* to *somewhat informative* prior distributions were defined by the study author, and represented a range in specificity for the prior distribution. For each of the three categories, (a) the means of the fixed effects from the previous sample's results was used as the prior distribution mean; and (b) the standard deviation ranged from the exact standard deviation for *very specific* prior distributions, two times the standard deviation for *specific* priors, and three times for *somewhat specific* priors. All prior distributions for fixed effects were specified as normally distributed.

Information criteria for contextual variables predicting fidelity of SWPBIS implementation growth are highlighted in Table 12. For the 2009-10 sample, all four model selection criteria suggested that the *informative* model was the best fitting. The AIC, BIC, and DIC were lowest for the *informative* model. The approximate Bayes factor comparing the *informative* model to the model with the next lowest BIC, the *somewhat informative* model, was 1.17. For the 2010-11 sample, three of four model selection criteria suggested that the *somewhat informative* model was the best fitting. The AIC and BIC were lowest for this model, and the approximate Bayes factor comparing the *somewhat informative* model to the *very informative* model was 3.57. For the 2011-12 sample, all four model selection criteria favored the *somewhat informative* model. The AIC, BIC, and DIC were lowest for this model, and the approximate Bayes

factor when comparing the *somewhat informative* model to the *very informative* model was 1.56. For the 2012-13 sample, the AIC, BIC, and DIC were lowest for the *very informative model*, and the approximate Bayes factor comparing this model to the *informative* model was 1.70. The results for models with converging model selection evidence are detailed in Table 13.

Table 12
Information Criteria for Multilevel Models with Contextual Covariates

	AIC	BIC	DIC
2008-09			
ML model	227.57	255.75	
Bayesian models			
Uninformative*	159.03	168.07	208.86
2009-10			
ML model	1252.08	1306.87	
Bayesian models			
Uninformative	1026.17	1059.94	1269.70
Somewhat informative	1025.15	1058.92	1252.24
Informative	1024.84	1058.61	1251.96
Very informative*	1030.71	1064.48	1291.69
2010-11			
ML model	1058.30	1108.84	
Bayesian models			
Uninformative*	973.91	1005.44	1142.83
Somewhat informative	961.60	993.13	1126.26
Informative	965.26	996.78	1122.63
Very informative	964.15	995.67	1141.00
2011-12			
ML model	1374.20	1428.57	
Bayesian models			
Uninformative*	1340.07	1374.61	1462.85
Somewhat informative	1185.98	1220.52	1355.38
Informative	1187.75	1222.29	1361.93
Very informative	1186.87	1221.41	1355.89
2012-13			
ML model	1750.48	1808.97	
Bayesian models			
Uninformative	1550.75	1589.83	1829.38
Somewhat informative	1549.06	1588.14	1819.79
Informative	1548.54	1587.62	1820.69
Very informative	1547.49	1586.57	1814.28

Note. *These models did not converge using the criterion of $\hat{R} < 1.1$ for all parameters.

The results for the 2012-13 analyses provided the most up to date estimate of contextual variables influence on fidelity growth and are interpreted. The intercept and slope, β_{00} and β_{10}, represent the estimate for schools with zero percent of students on free and reduced price lunch, have less than or equal to 500 students, and are located in cities. This is the case because of the variables included in the analysis are as follows: (a) The variable for percent of student on free and reduced price lunch is continuous with a range of 0.00 to 1.00; (b) Greater than 500 is an binary indicator for a variable representing a dichotomy of school size; and (c) Suburb, town, and rural are a series of dummy codes for a categorical locality variable representing a school's location in a city, suburb, town, or rural location. Given these considerations, the estimates for β_{00} and β_{10} represent estimates for a small group of schools. The posterior distribution mean for the intercept of this small group of schools was 22.91 with a 95% credibility interval of (21.04, 24.77). The posterior distribution mean for the slope of this small group of schools was 0.65 with a 95% credibility interval of (0.28, 1.02).

Based on results form the 2012-13 analysis, all contextual variables were influential predictors of the intercept. The posterior distribution mean for percent FRL was 4.85 with a 95% credibility interval of (2.13, 7.55). Given the FRL variable ranges from .00 to 1.00, the posterior distribution suggested that on average for every 10% gain in the percentage of students eligible for FRL, fidelity of SWPBIS implementation scores at time zero gained about half a point. The posterior distribution of greater than 500 students was 2.62 with a 95% credibility interval of (0.49, 4.77), implying that for schools with more than 500 students the fidelity of SWPBIS implementation score at time zero were expected to be half a point to almost 5 points higher than schools with less

than 500 students. The posterior distributions for the locality indicators suggested that schools suburbs, towns, and rural locations were expected to have lower fidelity scores at time zero than schools in cities. Schools in suburbs and towns were both expected to have fidelity scores on average 7.64 points lower at time zero than schools in cities with 95% credibility intervals of (-9.99, -5.31) and (-10.06, -5.24) respectively. Schools in rural areas were expected to have fidelity scores roughly three points lower at time zero than schools in cities with a 95% credibility interval of (-5.38, -0.66).

Table 13
Bayesian Multilevel Fidelity of Implementation Growth Models With Contextual Variables as Predictors and Updated Prior Distributions

	2008-09 (j = 13)	2009-10 (j = 61)	2010-11 (j = 53)	2011-12 (j = 64)	2012-13 (j = 85)
Fixed Effects					
Intercept: β_{00}	27.62 (16.4)	27.54 (0.2)	25.15 (1.0)	24.72 (0.5)	22.91 (0.9)
% FRL: β_{01}	7.14 (29.8)	7.14 (0.1)	7.02 (1.5)	6.97 (0.5)	4.85 (1.4)
> 500: β_{02}	7.96 (7.1)	7.86 (0.3)	5.34 (1.1)	4.88 (0.5)	2.62 (1.1)
Suburb: β_{03}	-11.07 (13.4)	-11.07 (0.2)	-10.36 (1.2)	-10.33 (0.5)	-7.64 (1.2)
Town: β_{04}	-9.13 (13.2)	-9.11 (0.2)	-8.48 (1.2)	-8.49 (0.5)	-7.64 (1.2)
Rural: β_{05}	-2.37 (12.6)	-2.44 (0.2)	-2.64 (1.2)	-2.83 (0.5)	-3.02 (1.2)
Time (slope): β_{10}	0.44 (2.7)	0.50 (0.1)	0.54 (0.2)	0.56 (0.2)	0.65 (0.2)
% FRL*time: β_{11}	;-1.15 (5.0)	-1.03 (0.3)	-0.66 (0.4)	-0.48 (0.3)	-0.65 (0.3)
> 500*time: β_{12}	-0.27 (1.2)	-0.64 (0.2)	-0.54 (0.3)	-0.33 (0.2)	-0.44 (0.2)
Suburb*time: β_{13}	-0.14 (2.1)	1.13 (0.2)	1.50 (0.3)	0.69 (0.2)	0.83 (0.2)
Town*time: β_{14}	0.72 (2.2)	1.02 (0.2)	0.99 (0.4)	1.00 (0.3)	0.87 (0.2)
Rural*time: β_{15}	0.48 (2.1)	0.06 (0.2)	0.44 (0.3)	0.39 (0.3)	0.46 (0.2)
Random Effects					
$\sigma_{\pi_{0j}}$	8.11 (2.3)	9.55 (0.9)	8.26 (1.0)	10.28 (1.0)	8.89 (0.8)
$\sigma_{\pi_{1j}}$	1.28 (0.4)	0.87 (0.1)	0.79 (0.2)	0.77 (0.1)	0.95 (0.1)
$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	-.82 (.2)	-.90 (.0)	-.88 (.1)	-.91 (.0)	-.94 (.0)
$\sigma_{e_{tj}}$	1.29 (0.2)	2.23 (0.2)	3.66 (0.3)	3.41 (0.2)	3.51 (0.2)
Intra-class correlations					
Level 1	.024	.051	.163	.098	.134
Level 2	.976	.949	.837	.902	.866

Note. For model values, the number outside the parentheses refers to the mean of the posterior distribution, and inside the parentheses is the posterior standard deviation. *j* refers to the number of schools in that year's sample.

The results of the 2012-13 analysis also suggested that all variables with the exception of the indicator for schools in a rural locations were influential predictors of the slope (i.e., the change in fidelity score for every month of school). The posterior distribution mean for the influence of the percentage of students on FRL was -0.65 with a 95% credibility interval of (-1.22, -0.06). This suggested that for every 10% increase of students on free and reduced priced lunch, the change in fidelity scores for each month of school could be expected to decrease by over half a point. The posterior distribution mean for the school size indicator's influence on the slope was -0.44 with a 95% credibility interval of (-0.78, -0.10), suggesting that schools with more than 500 students were expected to have about half a point less of fidelity growth per month of schools than schools with less than or equal to 500 students. The posterior distributions for indicators of schools in suburbs and towns suggested that on average schools could be expected to gain roughly 0.4 to 1.3 more fidelity points per month of school as compared to schools in cities, with 95% credibility intervals of (0.40, 1.25) and (0.40, 1.34) respectively. The posterior distribution for the indicator of schools in rural areas suggested that schools in rural areas had similar fidelity growth per month of school as schools in cities, with a 95% credibility interval of (-0.03, 0.94).

Again, it is important to look at the ML estimates as the Bayesian estimates take into account the likelihood distribution. As was the case with the years implementing predicting fidelity growth, the ML estimates result in different magnitudes and patterns of influence than the Bayesian estimates. As was discussed for the Bayesian estimates, all but one variable were influential predictors of initial status and growth fidelity growth parameters. Results of ML estimates highlighted in Table 14 show difference patterns of

Table 14

Maximum Likelihood Multilevel Fidelity of Implementation Growth Models With Contextual Variables as Predictors and Updated Prior Distributions

	2008-09 (j = 13)	2009-10 (j = 61)	2010-11 (j = 53)	2011-12 (j = 64)	2012-13 (j = 85)	Pooled (j = 276)
Fixed Effects						
Intercept: β_{00}	27.25	12.59	17.98	17.73	18.02	17.38
	(21.0)	(3.8)**	(2.6)***	(4.3)***	(2.6)***	(1.6)***
% FRL: β_{01}	4.77	14.89	6.35	-5.88	-2.73	3.17
	(22.4)	(6.0)**	(3.5)	(5.5)	(3.3)	(2.2)
> 500: β_{02}	7.83	1.68	-4.05	-1.95	0.82	-1.05
	(5.0)	(2.1)	(1.7)*	(2.2)	(1.7)	(1.0)
Suburb: β_{03}	-12.12	-3.31	0.68	3.30	3.56	1.01
	(9.4)	(3.1)	(2.0)	(3.6)	(2.0)	(1.3)
Town: β_{04}	-10.37	6.35	3.08	4.28	1.59	4.08
	(9.8)	(2.6)**	(2.5)	(4.3)	(2.5)	(1.5)**
Rural: β_{05}	-3.66	1.29	2.55	2.45	2.30	2.71
	(9.3)	(2.2)	(2.3)	(4.2)	(2.6)	(1.4)
Time (slope): β_{10}	0.36	1.68	1.11	1.10	1.24	1.28
	(3.3)	(0.5)***	(0.4)**	(0.4)*	(0.3)***	(0.2)***
% FRL*time: β_{11}	-0.58	-1.31	-0.55	0.53	0.13	-0.37
	(3.5)	(0.7)	(0.5)	(0.5)	(0.4)	(0.2)
> 500*time: β_{12}	-0.22	-0.13	0.40	0.20	-0.25	0.00
	(0.8)	(0.2)	(0.2)	(0.2)	(0.2)	(0.1)
Suburb*time: β_{13}	-0.01	0.75	0.48	-0.42	-0.37	-0.03
	(1.5)	(0.4)*	(0.3)	(0.3)	(0.3)	(0.1)
Town*time: β_{14}	0.90	-0.17	-0.04	-0.02	-0.09	-0.14
	(1.5)	(0.3)	(0.4)	(0.4)	(0.3)	(0.2)
Rural*time: β_{15}	0.74	-0.21	0.04	-0.10	-0.11	-0.15
	(1.4)	(0.3)	(0.3)	(0.4)	(0.3)	(0.2)
Random Effects						
$\sigma_{\pi_{0j}}$	5.78	6.54	4.62	8.12	6.57	7.26
$\sigma_{\pi_{1j}}$	0.87	0.65	0.33	0.56	0.68	0.66
$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	-.98	-.89	<-.99	-.97	-.94	-.90
$\sigma_{e_{ij}}$	1.29	2.23	3.74	3.47	3.55	3.24
Intra-class correlations						
Level 1	.047	.103	.394	.154	.224	.165
Level 2	.953	.897	.606	.846	.776	.835

Note. For model values, the number outside the parentheses refers to the maximum likelihood parameter estimate, and inside the parentheses is the standard error. j refers to the number of schools in that year's sample. *p < .05. **p < .01. ***p < .001.

influence. For example, even for the pooled model that takes into account all schools in all years, a model with arguably the most statistical power, only three fixed effect

parameters are significant: $\beta_{00,}$ $\beta_{10,\ and}$ β_{04}. The differences between the Bayesian and ML model results were due to the inclusion of the informative prior distribution in the Bayesian estimation algorithm.

Do sustainability factors predict fidelity growth? Introducing sustainability fixed effects into the model offered an interesting opportunity to model the Bayesian updating process with two samples. At the point of writing, only two samples were available with data from the sustainability measure, SUBSIST. Subsequently, the updating process was only employed once. Data for this analysis are based on a subsample of the 2010-11 involving ten schools. Demographic and descriptive information for these ten schools is provided in Tables 1 and 3. For this demonstration these ten schools will be considered the 2010-11 sample, but it should be noted that this different than the 2010-11 sample used in all other analyses. It is used here for demonstration purposes and limitations due to this inclusion procedure are discussed.

Demonstrating analyses sustainability factor effects on fidelity growth also presents the opportunity to show how to model growth with both level-two and level-three predictors. The models were specified according to Equations 9 and 10, and model code is available in Appendix G. Three levels were included in the model as sustainability predictors were both school and district level variables. Like the previous two demonstrations, models for this analysis were estimated using JAGS (Plummer, 2003) for the R program (R Core Team, 2012) and four Markov chains. The models were estimated using 100,000 iterations and a thinning rate of 80 to ensure facilitate model convergence. The models reported below had $\hat{R}$ values below 1.1 for all parameters indicating adequate model convergence. All models were specified using

normal distribution for fixed effect prior distributions, gamma distributions for level-1 error variance, and the Wishart distribution for level-2 and level-3 variance.

The information criteria for the Bayesian multilevel models are reported in Table 15. The 2010-11 uninformative posterior distributions were used as informative prior distributions for the 2012-13 analyses. Model selection criteria suggested that the *informative* model had the best fit for the 2012-13 sample. The AIC, BIC, and DIC were lowest for the *informative* model. The approximate Bayes factor comparing the *informative* model to the model with the next smallest BIC, the *somewhat informative* model, was 0.77 suggesting minimal evidence for the *informative* model over *somewhat informative* model. Given that three of four pieces of evidence point to the *informative* model, model results are interpreted below.

Table 15
Information Criteria for Multilevel Models with Sustainability Factors as Predictors

	AIC	BIC	DIC
2010-11			
ML model	178.84	203.21	
Bayesian model			
Uninformative	149.35	154.50	189.90
2012-13			
ML model	1718.27	1780.42	
Bayesian models			
Uninformative	1551.69	1593.22	1785.54
Somewhat informative	1546.43	1587.96	1789.30
Informative	1545.92	1587.44	1782.44
Very informative*	1551.30	1592.82	1792.93

Note. *This model did not converge using the criterion of $\hat{R} < 1.1$ for all parameters.

Table 16 documents model results for sustainability covariates predicting fidelity of implementation growth and results from the 2012-13 analysis are interpreted. The

sustainability predictors entered to the model as fixed effects at level-2 and level-3 level as they were school and district predictors of growth parameters. The posterior distribution mean of the intercept for all schools across all districts represents the average fidelity of implementation score at time zero for schools with factor scores of zero on sustainability factors, and was 27.57 with a 95% credibility interval of (26.82, 28.34). The posterior distribution mean of the slope for all schools across all districts represented the average fidelity growth during one month of school for schools with factor scores of zero on sustainability factors was 0.21 year with a 95% credibility interval of (-0.02, 0.44). Considering that the slope crossed zero suggested that the slope was not an influential predictor. Given that (a) the correlations between schools and districts for intercept and slopes were high and negative, and (b) the intercept is fairly high considering the maximum score on the fidelity measure is 34, a ceiling effect may be occurring. Subsequently, the slope for the analysis was non-influential for many schools and districts with relatively high intercepts, but is perhaps an important predictor for schools with lower intercepts.

The school variables were influential predictors of both the average initial fidelity scores and change in fidelity scores. For every one point increase on the school priority scale, the average initial fidelity score dropped by 22.91 points on average with a 95% credibility interval for the posterior distribution of (-23.32, -22.50), and the average change in fidelity score for every month of school increased by 2.11 with a 95% credibility interval of (1.49, 2.71). These results suggest that in schools with higher perceptions of priority for SWPBIS, the average initial measurement of SWPBIS fidelity was lower, but the rate of growth was higher than schools with lower school priority

levels. For every one point increase on the team use of data scale, the initial fidelity score increased on average 11.19 points with a 95% credibility interval for the posterior distribution of (10.68, 11.71), and the average change in fidelity score decreased by 0.98 points with a 95% credibility interval of (-1.34, -0.62). These results suggest that higher levels of perceived data use by SWPBIS teams were associated with higher initial fidelity levels and lower growth rates. Here a ceiling effect may be occurring because of the high intercept and large negative correlations between the slope and intercept. Descriptive statistics for this sample outlined in Table 3 show median for fidelity scores for the 2012-13 sample was 24 indicating that half the sample had fidelity scores of 24 or above. Further, 14.7% of observations (42 out of 286) from the 2012-13 sample had fidelity scores of 30 or above.

Both district variables were influential predictors of initial fidelity scores, and capacity building predicted change in fidelity scores. For every one point increase on the perceptions of district priority scale, the average initial fidelity score was 2.71 points higher with a 95% credibility interval for the posterior distribution of (2.29, 3.12), and the average change in fidelity score increased on average by 0.04 with a 95% credibility interval of (-0.55, 0.65). Because the posterior distribution for district priority variable's influence on the slope crossed zero, it can be concluded that district priority had relatively little influence on the change of fidelity scores during a school year. For every one point increase in the perceptions of capacity building scale, the average initial fidelity score was 10.42 points higher with a 95% credibility for the posterior distribution of (9.99, 10.83), and the average change in fidelity score decreased by 0.83 points with a 95% credibility interval of (-1.37, -0.29). Like the school predictor for perceptions of

team use of data, higher perception levels of district capacity building efforts were associated with higher initial fidelity levels and lower growth rates. Again, a ceiling effect may be occurring as schools approach the maximum score on the TIC.

Table 16
Bayesian Multilevel Fidelity of Implementation Growth Models With Sustainability Factors as Predictors and Updated 'Informative' Prior Distributions

	2010-11 (j = 10, k = 7)	2012-13 (j = 85, k = 45)
Fixed Effects		
Intercept: γ_{000}	27.98 (3.4)	27.57 (0.4)
School priority: β_{010}	-23.02 (11.6)	-22.91 (0.2)
Team use of data: β_{020}	11.08 (7.0)	11.19 (0.3)
District priority: γ_{001}	2.70 (11.4)	2.71 (0.2)
Capacity building: γ_{002}	10.42 (10.8)	10.42 (0.2)
Time (slope): γ_{100}	0.50 (0.5)	0.21 (0.1)
School priority*time: β_{110}	1.30 (1.6)	2.11 (0.3)
Team use of data* time: β_{120}	-0.76 (1.0)	-0.98 (0.2)
District priority*time: γ_{101}	0.35 (1.5)	0.04 (0.3)
Capacity building* time: γ_{102}	-0.92 (1.4)	-0.83 (0.3)
Random Effects		
$\sigma_{\pi 0j}$	4.24 (2.8)	6.87 (1.0)
$\sigma_{\pi 1j}$	0.50 (0.4)	0.63 (0.1)
$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	-.27 (.5)	-.92 (.1)
$\sigma_{\beta 00k}$	5.72 (3.1)	9.07 (1.7)
$\sigma_{\beta 10k}$	0.63 (0.4)	1.02 (0.2)
$\rho_{\sigma^2_{\beta 00k},\sigma^2_{\beta 10k}}$	-.50 (.4)	-.97 (.0)
$\sigma_{e_{tjk}}$	2.04 (0.4)	3.50 (0.2)
Intra-class correlations		
Level 1	.075	.086
Level 2	.329	.332
Level 3	.596	.582

Note. For model values, the number outside the parentheses refers to the mean of the posterior distribution, and inside the parentheses is the posterior standard deviation. j refers to the number of schools in that year's sample, and k refers to the number of districts.

As this example was based on two samples and only allowed one application of updates to prior distributions, it is also important to consider the influence of the likelihood and prior distributions relative to one another. To this end, the ML model estimates are detailed in Table 17. Looking closely at the ML results for 2012-13, several observations of differences as compared to the Bayesian models should be made. First, average intercept parameter, γ_{000}, is lower for the ML model and the average slope, γ_{100}, is higher and significant when compared to the Bayesian model. Second, the patterns of influence for school and district level predictors is different with team use of data being the only significant predictor of initial fidelity status and team use of data and capacity building significantly predicting fidelity growth. The magnitude of influence for these and other predictors was also different for the 2012-13 ML estimates when compared to the Bayesian estimates.

Referring back to the Bayesian estimates in Table 15, close examination of the 2010-11 estimates as compared to the 2012-13 estimates reveals that results are very similar. For example, the mean of the posterior distribution of the intercept in 2010-11 was 27.98 and 27.57 in 2012-13. Similarly, the mean of the posterior distribution of the fixed effect of team use of data on the initial status, β_{020}, in 2010-11 was 11.08 and 11.19 in 2012-13. The prior distribution's influence on these estimates accounts for these similarities. Also, the prior might be influencing poster results as the sample size for the 2012-13 sample in terms of number of schools, $j = 85$, and districts, $k = 45$, was still relatively small given that there were 17 parameters in the model. Given the influence of the prior, the analysis only included two samples, and the relatively small number of observations, caution is recommended before making inferences about the nature of

sustainability factors' influence on fidelity growth. In this case, collection of more samples is recommended.

Table 17
Maximum Likelihood Multilevel Fidelity of Implementation Growth Models With Sustainability Factors as Predictors

	2010-11 (j = 10, k = 7)	2012-13 (j = 85, k = 45)	Pooled (j = 95, k = 51)
Fixed Effects			
Intercept: γ_{000}	26.86 (2.5)***	21.05 (1.0)***	21.46 (0.9)***
School priority: β_{010}	-19.40 (4.1)***	1.52 (3.3)	-0.37 (3.2)
Team use of data: β_{020}	4.25 (2.5)	5.53 (2.3)*	5.83 (2.2)*
District priority: γ_{001}	6.94 (4.2)	4.43 (3.4)	5.62 (3.1)
Capacity building: γ_{002}	5.64 (4.0)	-4.94 (2.8)	-5.37 (2.7)
Time (slope): γ_{100}	0.57 (0.3)	0.83 (0.1)***	0.82 (0.1)***
School priority*time: β_{110}	1.25 (0.6)	0.00 (0.4)	0.13 (0.4)
Team use of data* time: β_{120}	-0.18 (0.4)	-0.66 (0.3)*	-0.67 (0.3)*
District priority*time: γ_{101}	-0.23 (0.5)	-0.53 (0.5)	-0.62 (0.4)
Capacity building* time: γ_{102}	-0.47 (0.6)	0.88 (0.4)*	0.90 (0.3)*
Random Effects			
$\sigma_{\pi 0j}$	0.00	3.98	4.42
$\sigma_{\pi 1j}$	0.00	0.25	0.28
$\rho_{\sigma^2_{\pi 0j},\sigma^2_{\pi 1j}}$	<-.99	-.95	-.96
$\sigma_{\beta 00k}$	6.42	4.52	4.63
$\sigma_{\beta 10k}$	0.60	0.61	0.61
$\rho_{\sigma^2_{\beta 00k},\sigma^2_{\beta 10k}}$	<-.99	<-.99	<-.99
$\sigma_{e_{tjk}}$	1.70	3.50	3.39
Intra-class correlations			
Level 1	.065	.250	.217
Level 2	.000	.326	.370
Level 3	.935	.424	.412

Note. For model values, the number outside the parentheses refers to the maximum likelihood parameter estimate, and inside the parentheses is the standard error. j refers to the number of schools in that year's sample, and k refers to the number of districts. *p < .05. *p < .01. ***p < .001.

CHAPTER V

DISCUSSION

Summary of Major Findings

This study provided a demonstration for modeling linear growth using a fully Bayesian approach incorporating information prior distributions for fixed effects in multilevel/hierarchical growth models. Three distinct growth models were demonstrated including (a) a two-level growth model with no predictors, (b) a two-level growth model with fixed-effect predictors at the second level, and (c) a three-level growth model with fixed-effect predictors at the second- and third-level. Further, a method for including informative prior distributions was outlined based on a serial updating process of prior distributions for fixed-effect predictors. Finally, a process for model selection was detailed. All of the above are embedded within the context of fidelity of implementation for a universal school systems intervention, School-Wide Positive Behavior Interventions and Supports (SWPBIS).

Results for from this demonstration are important for both methods research focused on Bayesian estimation and contextual SWPBIS research. Results from the demonstrations of a Bayesian approach to growth modeling showed that incorporation of informative prior distribution improved overall model fit as compared to models with uninformative priors distributions. Further, the model based inferences for informative Bayesian models were quite different than models estimated with uninformative priors and a maximum likelihood (ML) estimator. Determining whether informative Bayesian models provide more accurate representations of population parameters than models with uninformative priors or a ML estimator is beyond the scope of this demonstration, but

suggestions for simulations studies to explore this topic are outlined. In terms of SWPBIS research, the demonstrations included here documented the reliability of fidelity of implementation growth parameters, and provided preliminary evidence about predictors of this growth. The major findings, limitations, and suggestions for future research are further detailed in the following sections.

Bayesian Estimation. Results from this study have four main implications for researchers interested in Bayesian methodology. The four implications include (a) demonstrating a fully Bayesian approach to growth modeling, (b) how model selection criteria can be applied, (c) demonstrations documented better fit for models with informative priors, and (d) model results were different for Bayesian models with informative priors than ML models. These four topics are discussed in the following paragraphs.

First, this study provided a demonstration of a fully Bayesian approach to modeling multilevel linear growth. Specifically, methods for multilevel linear growth models using a Bayesian estimator were outlined. Further, methods for including fixed effects predictors at both the second- and third-level were detailed. Bayesian models explicitly incorporate probability into the reporting of results in the form of posterior distributions. In this sense, the results reported here were detailed in a Bayesian fashion with results in the form of distributions rather than point estimates. Considering that point estimates assume distributional form and have standard errors based on sampling theory, these types of tests introduce assumptions that may not be viable under all conditions (Kruschke, 2013). Multilevel ML models primary rely on *t*-tests to determine significance of fixed-effect parameters of interest. As the number of tests increase with

as more parameters are included in the model so does the threat of Type I error resulting in "statistically significant effects that are not in fact real" (Gelman et al., 2012, p. 190). In fact, Raudenbush and Bryk (2002, p. 283) recommend using a Bayes approach when making inferences based on fixed effects when analyzing small sample unbalanced data. Subsequently, a Bayesian approach to multilevel growth modeling may be preferred as results are probabilistic providing a more accurate representation of the uncertain nature of social science phenomenon such as fidelity growth detailed here within.

Second, this study documented how model selection criteria can be used in an objective Bayesian approach to statistical modeling. Gelman and Shalizi (2013) pointed out that Bayesian estimation was not an inductive endeavor, and that statistical modeling involved many decisions that in many instances can be objective. One way to make statistical modeling decisions objective is to explicitly incorporate numerical indices and selection criteria into the modeling process. To this end, this study showed how information criteria (specifically, AIC, BIC, and DIC) and approximate Bayes factors could be used to choose between competing models. Based on recommendations from Liu and Aitkin (2008), a process for model selection was employed based on converging evidence across four model selection indices including the AIC, BIC, DIC, and approximate Bayes factor. Models were chosen over competing models if multiple pieces of evidence favored one model over others. Further, all of the models favored in the demonstrations included had at least three of four criteria in favor of that model. In many cases, the four model selection criteria converged on a single model.

Based on model selection criteria, demonstrations provided evidence that models with informative prior distributions had better fit than models with uninformative prior

distributions. The model selection criteria detailed in this study included indices for both local and global model generalizability (Liu & Aitkin, 2008). Local generalizability refers to how well the model fits the data, and global generalizability refers to how well a model predicts future data. In all analyses that included comparing models with informative prior distributions to models with uninformative priors, the models selected based on model selection criteria had informative prior distributions. These findings align with Yu and Abdel-Aty's (2013) work that showed informative priors enhanced fit for multilevel models incorporating Poisson and gamma distributions of safety functions for accident prevention. This project extends the scope of their work, by showing that (a) models with informative priors not only fit better after two samples, but continued to exhibit better fit after five samples; and (b) documenting that informative priors enhanced model fit for models with normal distributions for fixed effects, gamma distributions for level-one variance, and Wishart distributions for level-two and –three random effects.

Fourth, model results for Bayesian estimates based on informative prior distributions were different than model results based solely on data distributions. Arguably, "the most controversial aspect of Bayesian statistics is the necessary assignment of a prior distribution" (Gill, 2009, p. 60). Given the results documented here from the use of informative prior distributions result in different model based inferences, the use of them in this study may be problematic. However, uninformative model results are nearly identical to results of ML models (Gelman & Hill, 2007; Gill, 2002), and, as already stated, the Bayesian models with informative priors outlined here resulted in more favorable model selection criteria than models with uninformative priors. Given these reasons, the Bayesian models with informative priors *may* be providing more

realistic estimates of SWPBIS fidelity growth and predictors of this growth. This conclusion is tenuous because of the scope of this study, and the relative influence of the prior distribution over observed data for parameter estimates, a topic addressed in the limitations.

Fidelity and School-Wide Positive Behavior Interventions and Support. The results of the demonstrations included in this study may be useful for researchers focused on intervention program fidelity and SWPBIS. Results documented reliability information for a within school-year fidelity growth metric based on repeated measures of the Team Implementation Checklist (Sugai, Todd, et al., 2001), a SWPBIS fidelity self-assessment. Further, results provided evidence for predictors of fidelity growth.

Based on reliability indices calculated for both the slope and intercept of the fidelity growth model, one thing of note was that the Bayesian models incorporating informative prior distributions were slightly more reliable than the ML models. For the Bayesian models, the intercept term had high reliability as evidenced by the correlation between the predicted and observed score at time zero, with indices ranging from .960 to .998 across five samples. The slope term's reliability fluctuated based on metric and sample. The correlation between observed and predicted change in fidelity scores was on the moderate to high end of the spectrum, with indices ranging from .676 to .978. The reliability indices based on Raudenbush and Bryk's (2002) reliability formula were low to moderate, ranging from .233 to .793, but were influenced by the relatively small sample size. These findings add to previous literature focused on rate of program implementation (Buzhardt et al., 2006) by providing preliminary psychometric evidence

for within year fidelity growth for schools during the installation phase of implementation (Fixsen et al., 2005).

Results of the first two demonstrations involving the use of informative prior distributions for fixed effects of predictors of SWPBIS fidelity growth parameters, revealed that the number of years implementing SWPBIS did not predict fidelity growth, but contextual variables such as the percent of students eligible for FRL, size, and locality did predict fidelity growth. Caution is recommended, however, as the analytic decision for centering of time limits the implications of these findings. Specifically, schools with higher percentages of students eligible for FRL and were larger than 500 students had higher intercept estimates and less steep slope estimates. Conversely, schools located in suburbs, towns, and rural areas had lower intercept estimates than schools in cities, and schools in suburbs and towns had steeper slope estimates than schools in cities. Schools in rural areas did not have statistically different slope estimates than schools in cities.

Results from a third demonstration of the influence of sustainability covariates on SWPBIS fidelity growth parameters using informative priors showed that school and district sustainability factors predicted fidelity growth. Caution is recommended interpreting these results, as they were collected from only two samples, and again the choice of centering time limits the implications of this finding. Preliminary evidence showed that (a) the school-level variables for 'school priority' and 'team use of data' were influential predictors of the intercept and slope; and (b) the district-level variables for 'district priority' and 'capacity building' were influential predictors of the intercept, and 'capacity building' was an influential predictor of slope.

Limitations

Several limitations should be noted. First, several limitations exist due to the samples included in the study that have implications for Bayesian methodology and SWPBIS research. Second, limitations arose because of the outcome fidelity measure used. Third, challenges due to lack of convergence for several of the Bayesian models will be discussed. Fourth, muliticolinearity of fidelity growth predictors might have biased results. And finally, limitations arose due to the choice on centering time.

The samples included in this study limit the generalizability of results to all schools implementing SWPBIS, and potentially limit implications for Bayesian methodology. All of the samples included in this study were convenience samples of schools using an online SWPBIS database (i.e., the OSEP Technical Assistance Center on Effective Schoolwide Interventions: Positive Behavioral Interventions and Supports, www.pbis.org). Further, the schools also self-elected to use the TIC at least three times per school year to monitor implementation fidelity, and met inclusion criteria. Subsequently, the schools that comprised the five samples could be functionally different than other schools not included in the samples. Additionally, while all samples were on the small size, the 2008-09 sample was notably smaller than the other samples ($j = 13$). Further, descriptive statistics of the 2008-09 sample show that it has the highest mean of outcome variable when compared to the other four samples. Subsequently, posterior distribution estimates that were serially applied as prior distributions might be overly influential on posterior distributions of later samples.

Limitations also arose due to the use of the outcome measure. Given that the measure used for this study is comprised of the 17 consistent items across three versions

of the TIC, the reliability information does not generalize to any one measure. Further, given that only limited information is known about the validity of the TIC (Tobin et al., 2012), the generalizability of findings to SWPBIS fidelity research is limited.

As several Bayesian models did not completely converge, the results may be biased. For the demonstrations including predictors of the fidelity growth, models were estimated using more MCMC iterations, yet several models failed to adequately converge. Given the nature of the study was to demonstrate the use of Bayesian methods for multilevel growth modeling, convergence failures were included in the reporting of results to highlight that this can occur. For the models with indicated convergence issues, the most common convergence failure was for the correlations between random effect growth parameters and other model parameters appeared to converge adequately.

Contextual variables predicting SWPBIS fidelity growth might have been multicolinear and results may reflect these associations of predictor variables. It was beyond the scope of this study to fully tease out the associations of predictor variables, and these relations may be influencing the results reported. For example, large schools might be located in city settings more frequently, implying a correlation between these two factors. Associations such as these were not explored in the models and might have biased the results.

Finally, a major limitation arose because of the choice for centering time. To fully facilitate the methods demonstrate, I chose to center time for each sample at the first measurement occasion for each school. This decision created a time variable where values of zero reflected different times of the school year for each school in each sample. The interpretation of the intercept was different for each school because some schools

completed their first fidelity measure towards the beginning of the school year, while others completed their first fidelity measure at the middle of the school year. The choice to center time at the initial fidelity measurement occasion for each school in each sample was driven by the desire to have enough information to accurately estimate intercept parameters enabling the demonstration of the modeling approach. Consequently, results pertaining to the intercept parameters lack substantive interpretations and results of slope parameters may not accurately reflect the fidelity growth phenomenon.

More specifically, the fidelity growth results may not represent an accurate depiction of the fidelity growth phenomenon limiting the findings implications for SWPBIS research. Choice of centering time implied that the intercept represented a fidelity score at the first measurement occasion for a particular school, but did not represent an expected fidelity score for any one point in the implementation process. As models were estimated using multilevel regression techniques, each schools intercept was the expected value for its first fidelity measurement occasion and the average intercept across all schools did not represent an average fidelity measure at any particular moment in the implementation process. Choice of centering time may also have implications for the slope estimates as choice on centering time has been shown to impact the variance of the intercept and covariance between the intercept and slope (Mehta & West, 2000).

Implications for Future Research

Both the findings and limitations of this study provide several avenues for future research on Bayesian methodology. Given the finding that Bayesian models with informative priors had overall better model fit than uninformative models, the extent to which this is the case in other contexts and simulations studies should be explored.

Further, the extent that model based inferences for Bayesian models with informative priors based on the serial updating process generalize to the *true* parameters could be explored with simulation studies. The extent to which Bayesian results based on serially updated to prior distributions compare to uninformative and ML results and which models provide a better understanding of *true* parameter values should also be explored through simulation studies. Lastly, based on the influence of the first sample results on the four subsequent sample posterior estimates documented in this study, the extent that *overly influential* samples influence the Bayesian updating process could be explored through simulation studies.

These ideas for future simulations could be bundled into factors for simulation studies to better understand how well linear growth models estimate known population parameters when varying sampling and analytic conditions. The basic process of conducting a Monte Carlo simulation experiment involves six steps including: Stating the research problem, specifying a the experimental plan, simulating data, estimating the statistical model being examined, replicating the process, and analyzing results from replications (Skrondal, 2000). The discussion above alludes to several factors that could be varied in a simulation study including estimator, sample size, number of samples included in each analysis applying BU (i.e., two samples, five, ten, etc.), the magnitude of influence for *overly influential* samples included in a BU analysis (i.e., moderately or largely different from other samples), and where in the BU process is the *overly influential* sample placed (i.e., it is the first sample analyzed, somewhere in the middle, at the end, etc.). If all of these factors were included, the resulting simulation study would include five factors with anywhere from two to four or more levels per factor, resulting in

a large fully crossed simulation design. Subsequently, it may be more feasible to explore two or three of these factors in relation to one another.

The basic question of whether it is worth employing a BU analysis or simply pooling samples for simultaneous analysis could be explored with a refined simulation study. For example, it may be valuable to compare the sample size and number of samples included in a series used for a BU analysis between two estimation conditions. The first estimation condition could involve serially applying a BU estimation process where results from the last sample in the serial represent the final statistical estimate. These results could be compared to results from a Bayesian model with an uninformative prior distribution where all observations across samples in the same series are pooled into one sample. The experiment could be further varied to enhance generalizability to other analytic conditions by: (a) Varying the sample size for each sample in a series of samples with factor levels of 20, 50, and 100; and (b) varying the number of samples per series with factor levels of five, ten, and 20. This example would result in a two by three by three design resulting in 18 unique conditions. Based on recommendations from Skrondal (2000), an analysis of variance meta-model could be specified to explore the main and two-way interaction effects of varying analytic conditions based on simulation study procedures outlined in the previous paragraph (where the number of observations for the meta-model is the number of replications). This method could be used to compare the precision of linear growth estimates produced through a BU process to that of a Bayesian estimate with uninformative priors and pooled samples.

The findings of the demonstrations provided avenues for future SWPBIS research. The evidence here provided preliminary evidence on both the reliability of fidelity

growth and possible predictors of that growth. Future research could focus on the documenting the validity of the TIC instrument to provide evidence for the generalizability of fidelity growth. To fully understand the fidelity growth of SWPBIS and related predictors, choice of centering the time variable could be altered to better address substantive questions (Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004) related to the implementation process. Also, further research should continue to explore the extent to which predictors included here and additional variables influence fidelity growth as this information could aid practitioners implement SWPBIS with more fidelity. Also, linking fidelity growth and valued outcomes such as office discipline referral rates and academic performance could add to a growing literature base linking the implementation process and student-level variables (e.g., Horner et al., 2009). Finally, exploration of SWPBIS sustainability factors influence on fidelity growth should be continued.

Conclusion

This study demonstrated the use of Bayesian updating to form informative prior distributions for multilevel linear growth models. Results suggest that informative priors may enhance model fit, but more research is needed to validate this finding. As social science focuses on making probabilistic statements about phenomena of interest and the growing availability of extant data provides opportunities to analyze data collected across samples, new techniques such as the one outlined should be explored for their viability.

APPENDIX A

TEAM IMPLEMENTATION CHECKLIST (TIC) VERSION 3.1 (SUGAI, HORNER, LEWIS-PALMER, & ROSSETTO DICKEY, 2011)

PBIS Team Implementation Checklist (TIC 3.1)

This checklist is designed to be completed by the PBIS Team once a quarter to monitor activities for implementation of PBIS in a school. The team should complete the **Action Plan** at the same time to track items that are In Progress or Not Yet Started items.

School: Coach: Date of Report:

District: County: State:

Person Completing Report:

PBIS Team Members:

Complete & submit to coach quarterly. Status: **A** = **A**chieved, **I** = **I**n Progress, **N** = **N**ot Yet Started					
Date:					
ESTABLISH COMMITMENT					
1. Administrator's Support & Active Involvement • Admin attends PBIS meetings 80 % of time • Admin defines social behavior as one of the top three goals for the school • Admin actively participates in PBIS training	Status:				
2. Faculty/Staff Support • 80% of faculty document support that school climate/discipline is one of top three school improvement goals • Admin/faculty commit to PBIS for at least 3 years	Status:				
ESTABLISH & MAINTAIN TEAM					
3. Team Established (Representative) • Includes grade level teachers, specialists, paraprofessionals, parents, special educators, counselors. • Team has established clear mission/purpose	Status:				
4. Team has regular meeting schedule, effective operating procedures • Agenda and meeting minutes are used • Team decisions are identified, and action plan developed	Status:				
5. Audit is completed for efficient integration of team with other teams/initiatives addressing behavior support • Team has completed the "Working Smarter" matrix	Status:				
Complete & submit to coach quarterly. **Status: A** = **A**chieved, **I** = **I**n Progress, **N** = **N**ot Yet Started					
Date:					

Complete quarterly with your PBIS Coach

Team Implementation Checklist, Version 3.1, September, 2011
Sugai, G., Horner, R., Lewis-Palmer, T., & Rossetto Dickey, C.
Adapted from Sugai, Horner, Lewis-Palmer, 2001
Educational and Community Supports, University of Oregon

SELF-ASSESSMENT					
6. Team completes self-assessment of current PBIS practices being used in the school • The staff completes the TIC (progress monitoring), BoQ (annual assessment) or SET.	Status				
7. Team summarizes existing school discipline data • The team uses office discipline referral data (ODR), attendance, & other behavioral data for decision making.	Status				
8. Team uses self-assessment information to build implementation Action Plan (areas of immediate focus) • The team uses the Action Plan to guide PBIS implementation.	Status:				
ESTABLISH SCHOOL-WIDE EXPECTATIONS: PREVENTION SYSTEMS					
9. 3-5 school-wide behavior expectations are defined and posted in all areas of building • 3-5 positively and clearly stated expectations are defined. • The expectations are posted in public areas of the school.	Status:				
10. School-wide teaching matrix developed • Teaching matrix used to define how school-wide expectations apply to specific school locations. • Teaching matrix distributed to all staff.	Status:				
11. Teaching plans for school-wide expectations are developed • Lesson plans developed for teaching school-wide expectations at key locations throughout the school. • Faculty is involved in development of lesson plans.	Status:				
12. School-wide behavioral expectations taught directly & formally • Schedule/plans for teaching the staff the lessons plans for students are developed • Staff and students know the defined expectations. • School-wide expectations taught to all students • Plan developed for teaching expectations to students to who enter the school mid-year.	Status:				

Complete & submit to coach quarterly. **Status: A** = **A**chieved, **I** = **I**n Progress, **N** = **N**ot Yet Started				
Date:				

Complete quarterly with your PBIS Coach

Team Implementation Checklist, Version 3.1, September, 2011
Sugai, G., Horner, R., Lewis-Palmer, T., & Rossetto Dickey, C.
Adapted from Sugai, Horner, Lewis-Palmer, 2001
Educational and Community Supports, University of Oregon

13. System in place to acknowledge/reward school-wide expectations • Reward systems are used to acknowledge school-wide behavioral expectations. • Ratio of reinforcements to corrections is high (4:1). • Students and staff know about the acknowledgement system & students are receiving positive acknowledgements.	Status:				
14. Clearly defined & consistent consequences and procedures for undesirable behaviors are developed • Major & minor problem behaviors are all clearly defined. • Clearly defined and consistent consequences and procedures for undesirable behaviors are developed and used. • Procedures define an array of appropriate responses to minor (classroom managed behaviors). • Procedures define an array of appropriate responses to major (office managed) behaviors.	Status:				
CLASSROOM BEHAVIOR SUPPORT SYSTEMS					
15. School has completed a school-wide classroom systems summary • The teaching staff has completed a classroom assessment (Examples: SAS Classroom Survey, Classroom Systems Survey, etc.)	Status:				
16. Action plan in place to address any classroom systems identified as a high priority for change • Results of the assessment are used to plan staff professional development and support.	Status:				
17. Data system in place to monitor office discipline referral rates that come from classrooms • School has a way to review ODR data from classrooms to use in data based decision making.	Status:				
18. Discipline data are gathered, summarized, & reported at least quarterly to whole faculty • Data collection is easy, efficient & relevant for decision-making • ODR data entered at least weekly (min). • Office referral form lists a) student/grade, b) date/time, c) referring staff, d) problem behavior, e) location, f) persons involved, g) probable motivation, h) consequences and i) administrative decision. • ODR data are available by frequency, location, time, type of problem behavior, motivation and student. • ODR data summary shared with faculty at least monthly (min).	Status:				

Complete quarterly with your PBIS Coach

Complete & submit to coach quarterly. **Status: A = Achieved, I = In Progress, N = Not Yet Started**					
	Date:				
19. Discipline data are available to the Team regularly (at least monthly) in a form and depth needed for problem solving • Team is able to use the data for decision making, problem solving, action planning and evaluation. • Precision problem statements are used for problem solving.	Status:				
BUILD CAPACITY FOR FUNCTION-BASED SUPPORT					
20. Personnel with behavioral expertise are identified & involved • Personnel are able to provide behavior expertise for students needing Tier II and Tier III support.	Status:				
21. At least one staff member of the school is able to conduct simple functional behavioral assessments • At least one staff member can conduct simple behavioral assessments and work with a team in developing behavior support plans for individual students	Status:				
22. Intensive, individual student support team structure in place to use function-based supports • A team exists that focuses on intensive individualized supports for students needing Tier III supports. • The team uses function-based supports to develop, monitor and evaluate behavioral plans. • The team delivering Tier III has a data system that allows on-going monitoring of the fidelity and outcomes of individual behavior support plans.	Status:				

Additional Comments & Information:

Complete quarterly with your PBIS Coach

Team Implementation Checklist, Version 3.1, September, 2011
Sugai, G., Horner, R., Lewis-Palmer, T., & Rossetto Dickey, C.
Adapted from Sugai, Horner, Lewis-Palmer, 2001
Educational and Community Supports, University of Oregon

APPENDIX B

SCHOOL-WIDE UNIVERSAL BEHAVIOR SUSTAINABILITY INDEX: SCHOOL TEAMS (SUBSIST) (MCINTOSH, DOOLITTLE, VINCENT, HORNER, & ERVIN, 2009)

Each page includes a number of statements (for example, 1.1. SW-PBIS (aka School-wide PBS, PBIS, EBS) serves a critical need for the school).

For each statement, you will be asked whether the statement is true for your school right now.

1.1. SW-PBIS (aka School-wide PBS, PBIS, EBS) serves a critical need for the school.

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.2. SW-PBIS addresses outcomes that are highly valued by school personnel.

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.3. A vast majority of school personnel (80% or more) support SW-PBIS.

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.4. SW-PBIS has been integrated into new school or district initiatives (e.g., renamed to meet new needs, shown how it can meet the goals of the new initiatives as well).

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.5. Parents are actively involved in the SW-PBIS effort (e.g., as part of SW-PBIS team or district committee)

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.6. The school administrators describe SW-PBIS as a top priority for the school.

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.7. The school administrators actively support school personnel when implementing and aligning initiatives (e.g., shield staff from competing demands, change language to align SW-PBIS with new initiatives) to allow SW-PBIS to occur.

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.8. A school administrator regularly attends and participates in SW-PBIS team meetings.

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.9. The practices and strategies of SW-PBIS are evidence-based (i.e., there is published research documenting their effectiveness).

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.10. School personnel perceive SW-PBIS as effective in helping them achieve desired outcomes.

Not True | Partially true | Mostly true | Very true | Don't know

1.11. School personnel celebrate the positive effects of SW-PBIS at least yearly.

Not true | Partially true | Mostly true | Very true | Don't know/NA

1.12. SW-PBIS has a "crossover effect" in other areas (e.g., improved academic achievement scores, attendance).

Not true | Partially true | Mostly true | Very true | Don't Know/NA

1.13. SW-PBIS is effective for a large proportion of students.

Not true | Partially true | Mostly true | Very true | Don't Know/NA

1.14. SW-PBIS has been expanded to other areas (e.g., classrooms, buses, students with intensive needs, parenting workshops).

Not true | Partially true | Mostly true | Very true | Don't Know/NA

1.15. SW-PBIS is implemented with fidelity (i.e., it is used as intended).

Not true | Partially true | Mostly true | Very true | Don't Know/NA

1.16. SW-PBIS becomes easier to use with continued experience.

Not true | Partially true | Mostly true | Very true | Don't Know/NA

1.17. SW-PBIS is considered to be a typical operating procedure of the school (it has become "what we do here/what we've always done")

Not true | Partially true | Mostly true | Very true | Don't Know/NA

1.18. SW-PBIS is cost-effective (in terms of money and effort).

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

1.19. Data collected for SW-PBIS are easy to collect and do not interfere with teaching.

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

1.20. Materials related to SW-PBIS (e.g., handbook, posters) can be used or adapted with ease across years.

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

1.21. There is an immediate (within 6 months) effect of SW-PBIS (e.g., reduction in referrals/suspensions, improved school climate, improved student success) after implementation.

Not true	Partially true	Mostly true	Very true	Don't know/NA
○	○	○	○	○

2.1. The school team implementing SW-PBIS is knowledgeable and skilled in SW-PBIS.

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

2.2. The school team implementing SW-PBIS is well organized and operates efficiently.

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

2.3. The school team implementing SW-PBIS meets at least monthly.

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

2.4. Needs assessments (e.g., EBS/PBIS Self Assessment Survey) are conducted.

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

2.5. There is regular measurement of fidelity of implementation (e.g., Team Implementation Checklist, School-wide Evaluation Tool, Benchmarks of Quality).

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

2.6. There is regular measurement of student outcomes (e.g., ODRs, achievement data, school safety surveys, student/parent satisfaction surveys).

Not true | Partially true | Mostly true | Very true | Don't Know/NA

2.7. Data are reviewed regularly at team meetings.

Not true | Partially true | Mostly true | Very true | Don't Know/NA

2.8. Data are presented to all school personnel at least four times per year.

Not true | Partially true | Mostly true | Very true | Don't Know/NA

2.9. Data are presented at least once per year to key stakeholders outside of the school (e.g., district officials, school boards, community agencies/groups).

Not true | Partially true | Mostly true | Very true | Don't Know/NA

2.10. Data are used for problem solving, decision making, and action planning (to make SW-PBIS more effective &/or efficient).

Not true | Partially true | Mostly true | Very true | Don't Know/NA

2.11. All school personnel have a basic understanding of SW-PBIS (i.e., know the critical features and practices).

.

Not true | Partially true | Mostly true | Very true | Don't Know/NA

D1.1. There are adequate district resources (funding and time) allocated for SW-PBIS.

Not true | Partially true | Mostly true | Very true | Don't know/NA

D1.2. The district administration actively supports SW-PBIS (e.g., describes SW-PBIS as a top priority, provides clear direction).

Not true | Partially true | Mostly true | Very true | Don't know/NA

D1.3. State/provincial officials actively support SW-PBIS (e.g., promotion, publicity, providing infrastructure).

Not true | Partially true | Mostly true | Very true | Don't know/NA

D1.4. SW-PBIS is promoted and visible to important organizations (e.g., school board, community agencies, businesses, parent groups).

Not true | Partially true | Mostly true | Very true | Don't know/NA

D1.5. SW-PBIS is embedded into school and/or district policy (e.g., school improvement plans, mission/vision statements).

Not true | Partially true | Mostly true | Very true | Don't know/NA

D2.1. The school team has regular access to district SW-PBIS expertise (e.g., external/district coaches or consultants).

Not true | Partially true | Mostly true | Very true | Don't Know/NA

D2.2. School teams and new personnel are provided with professional development in SW-PBIS at least yearly.

Not true | Partially true | Mostly true | Very true | Don't Know/NA

D2.3. The school team is connected to a "community of practice" (e.g., network of other SW-PBIS schools in district, local/regional conferences).

Not true | Partially true | Mostly true | Very true | Don't Know/NA

B1.1. School personnel are opposed to SW-PBIS because it goes against their personal values (e.g., "rewarding" students, teaching "compliance").

Not true | Partially true | Mostly true | Very true | Don't Know/NA

B1.2. Other school/district initiatives (e.g., academic, behavior) are present that compete (for time, resources or content) with SW-PBIS.

Not true | Partially true | Mostly true | Very true | Don't Know/NA

B1.3. There are high levels of turnover of school administrators (i.e., yearly).

Not true | Partially true | Mostly true | Very true | Don't Know/NA

B1.4. There are high levels of turnover of school personnel who served as key leaders ("champions") of SW-PBIS (i.e., within three years).

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

B1.5. There are high levels of general school personnel turnover (i.e., 50% of staff).

Not true	Partially true	Mostly true	Very true	Don't Know/NA
○	○	○	○	○

How often does your school SW-PBIS team currently meet (during the school year)?

Weekly	Every other week	Monthly	Every 6 weeks	Every other month	Other (please specify):
○	○	○	○	○	○

How often are data presented to all school personnel?

Weekly	Every other week	Monthly	Every 6 weeks	Every other month	4 times per year	3 times per year	2 times per year	once per year	less than once per year
○	○	○	○	○	○	○	○	○	○

Does this school have an external coach/facilitator/consultant with official work hours (FTE) dedicated to supporting SW-PBIS?

Yes	No
○	○

If you would like to receive a gift card for participation, please enter your name and street address where you would like it sent. If you would like your answers to remain anonymous, please leave the boxes blank.

Name

Street Address (including City, State, & zip)

Last year, someone at your school (most likely you!) tallied and provided information regarding the number of trainings attended, coaching access, and peer networking events for the year. Would you like to complete the ADEPT (coaching and training log) again this year for an additional $50 gift card?

Yes	No
○	○

If so, please provide your e-mail address here:

If not, please suggest other (e.g., coach, other team members) who may be interested in tallying coaching and

training for this year. Please provide their names and emails:

APPENDIX C

TWO LEVEL GROWTH MODEL CODE

```
model <- function(){
  for(t in 1:K){
    y[t] ~ dnorm(y.hat[t], tau.y)
    y.hat[t] <- pi0[SchId[t]] + pi1[SchId[t]]*Time[t]
  }
  sigma.y <- pow(tau.y, -0.5)
  tau.y ~ dgamma(1, 1) #prior level 1 variance

  for(j in 1:J){
    pi0[j] <- xi.pi0*Pi.raw[j,1]
    pi1[j] <- xi.pi1*Pi.raw[j,2]
    Pi.raw[j,1:2] ~ dmnorm(Pi.raw.hat[j,], Tau.Pi.raw[,])
    Pi.raw.hat[j,1] <- b00.raw
    Pi.raw.hat[j,2] <- b10.raw
  }
  b00.raw <- b00/xi.pi0
  b10.raw <- b10/xi.pi1
  b00 ~ dnorm(b00.mean, b00.sd) #level 2 prior
  b10 ~ dnorm(b10.mean, b10.sd) #level 2 prior

  xi.pi0 ~ dunif(0, 100)
  xi.pi1 ~ dunif(0, 100)

  Tau.Pi.raw[1:2,1:2] ~ dwish(W[,], df)
  df <- 3
  Sigma.Pi.raw <- inverse(Tau.Pi.raw)
  sigma.pi0 <- xi.pi0*sqrt(Sigma.Pi.raw[1,1])
  sigma.pi1 <- xi.pi1*sqrt(Sigma.Pi.raw[2,2])
  rho <- Sigma.Pi.raw[1,2]/sqrt(Sigma.Pi.raw[1,1]*Sigma.Pi.raw[2,2])
}
```

APPENDIX D

TWO LEVEL GROWTH MODEL WITH PREDICTORS CODE

```
model <- function(){
  for(t in 1:N){
    y[t] ~ dnorm(y.hat[t], tau.y)
    y.hat[t] <- inprod(Pi[SchId[t],], Time[t,])
  }
  sigma.y <- pow(tau.y, -0.5)
  tau.y ~ dgamma(1, 1)

  for(k in 1:K){
    for(j in 1:J){
      Pi[j,k] <- xi[k]*Pi.raw[j,k]
    }
    xi[k] ~ dunif(0, 100)
  }
  for(j in 1:J){
    Pi.raw[j,1:K] ~ dmnorm(Pi.raw.hat[j,], Tau.Pi.raw[,])
    for(k in 1:K){
      Pi.raw.hat[j,k] <- inprod(B.raw[k,], U[j,])
    }
  }
  for(k in 1:K){
    for(l in 1:L){
      B.raw[k,l] <- B[k,l]/xi[k]
      B[k,l] ~ dnorm(B.prior.mean[k,l], B.prior.sd[k,l])
    }
  }

  Tau.Pi.raw[1:2,1:2] ~ dwish(W[,], df)
  df <- 3
  Sigma.Pi.raw <- inverse(Tau.Pi.raw)
  sigma.pi0 <- xi[1]*sqrt(Sigma.Pi.raw[1,1])
  sigma.pi1 <- xi[2]*sqrt(Sigma.Pi.raw[2,2])
  rho <- Sigma.Pi.raw[1,2]/sqrt(Sigma.Pi.raw[1,1]*Sigma.Pi.raw[2,2])
}
```

APPENDIX E

THREE LEVEL GROWTH MODEL WITH PREDICTORS CODE

```
model<- function(){
  for(t in 1:N){ #N is total # of observations
    y[t] ~ dnorm(y.hat[t], tau.y)
    y.hat[t] <- inprod(Pi[SchId[t],], Time[t,])
  }
  sigma.y <- pow(tau.y, -0.5)
  tau.y ~ dgamma(1, 1) #prior level 1 variance

  for(k in 1:K){
    for(j in 1:J){
      Pi[j,k] <- xi[k]*Pi.raw[j,k]
    }
    xi[k] ~ dunif(0, 100)
  }
  for(j in 1:J){
    Pi.raw[j,1:K] ~ dmnorm(Pi.raw.hat[j,], Tau.Pi.raw[,])
    for(k in 1:K){
      Pi.raw.hat[j,k] <- B.int.raw[DistId[j],k] + inprod(B.raw[k,], X[j,])
    }
  }
  for(k in 1:K){
    for(l in 1:L){
      B.raw[k,l] <- B[k,l]/xi[k]
      B[k,l] ~ dnorm(B.prior.mean[k,l], B.prior.sd[k,l])
    }
    for(m in 1:M){
      B.int[m,k] <- xi[k]*B.int.raw[m,k]
    }
  }
  for(m in 1:M){
    B.int.raw[m,1:K] ~ dmnorm(B.int.raw.hat[m,], Tau.B.int.raw[,])
    for(k in 1:K){
      B.int.raw.hat[m,k] <- inprod(G.raw[k,], U[m,])
    }
  }
```

(continued on next page)

```
  for(k in 1:K){
    for(p in 1:P){
      G.raw[k,p] <- G[k,p]/xi[k]
      G[k,p] ~ dnorm(G.prior.mean[k,p], G.prior.sd[k,p])
    }
  }

  Tau.Pi.raw[1:2,1:2] ~ dwish(W1[,], df1)
  df1 <- 3
  Sigma.Pi.raw <- inverse(Tau.Pi.raw)
  sigma.pi0 <- xi[1]*sqrt(Sigma.Pi.raw[1,1])
  sigma.pi1 <- xi[2]*sqrt(Sigma.Pi.raw[2,2])
  rho.pi <- Sigma.Pi.raw[1,2]/sqrt(Sigma.Pi.raw[1,1]*Sigma.Pi.raw[2,2])

  Tau.B.int.raw[1:2,1:2] ~ dwish(W2[,], df2)
  df2 <- 3
  Sigma.B.raw <- inverse(Tau.B.int.raw)
  sigma.b00 <- xi[1]*sqrt(Sigma.B.raw[1,1])
  sigma.b10 <- xi[2]*sqrt(Sigma.B.raw[2,2])
  rho.b <- Sigma.B.raw[1,2]/sqrt(Sigma.B.raw[1,1]*Sigma.B.raw[2,2])
}
```

REFERENCES CITED

Adelman, H.S., & Taylor, L. (2003). On sustainability of project innovations as systemic change. *Journal of Educational & Psychological Consultation, 14*(1), 1-25. doi: 10.1207/S1532768XJEPC1401_01

Algozzine, K., & Algozzine, B. (2007). Classroom instructional ecology and school-wide positive behavior support. *Journal of Applied School Psychology, 24*(1), 29-47. doi: 10.1300/J370v24n01_02

Almond, R.G., Mulder, J., Hemat, L.A., & Yan, D. (2009). Bayesian Network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics, 34*, 491-521. doi: 10.3102/1076998609332751

Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology, 66*, 1-7.

Baker, S., Gersten, R., Dimino, J.A., & Griffiths, R. (2004). The sustained use of research-based instructional practice: A case study of Peer-Assisted Learning Strategies in mathematics. *Remedial and Special Education, 25*(1), 5-24. doi: 10.1177/07419325040250010301

Bambara, L.M., Nonnemacher, S., & Kern, L. (2009). Sustaining school-based individualized positive behavior support: Perceived barriers and enablers. *Journal of Positive Behavior Interventions, 11*(3), 161-176. doi: 10.1177/1098300708330878

Bayes, T., & Price, R. (1763). An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions (1683-1775)*, 370-418.

Benz, M.R., Lindstrom, L., Unruh, D., & Waintrup, M. (2004). Sustaining secondary transition programs in local schools. *Remedial and Special Education, 25*(1), 39-50.

Berman, P., & McLaughlin, M.W. (1976). Implementation of educational innovation. *Educational Forum, 40*, 344-370.

Biancarosa, G., Bryk, A.S., & Dexter, E.R. (2010). Assessing the value-added effects of Literacy Collaborative professional development on student learning. *The Elementary School Journal, 111*(1), 7-34. doi: 10.1086/653468

Biesanz, J.C., Deeb-Sossa, N., Papadakis, A.A., Bollen, K.A., & Curran, P.J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods, 9*(1), 30-52.

Bolfarine, H., & Bazan, J.L. (2010). Bayesian estimation of the logistic Positive exponent IRT model. *Journal of Educational and Behavioral Statistics, 35*(6), 693-713. doi: 10.3102/1076998610375834

Botzen, W.J.W., & van den Bergh, J.C.J.M. (2012). Risk attitudes to low-probability climate change risks: WTP for flood insurance. *Journal of Economic Behavior & Organization, 82*(1), 151-166. doi: 10.1016/j.jebo.2012.01.005

Bradshaw, C.P., Koth, C.W., Bevans, K.B., Ialongo, N., & Leaf, P.J. (2008). The impact of school-wide positive behavioral interventions and supports (PBIS) on the organizational health of elementary schools. *School Psychology Quarterly, 23*(4), 462-473. doi: 10.1037/a0012883

Bradshaw, C.P., Mitchell, M.M., & Leaf, P.J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions, 12*(3), 133-148. doi: 10.1177/1098300709334798

Bryk, A.S., & Raudenbush, S.W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*(1), 65-108. doi: 10.1086/443913

Buzhardt, J., Greenwood, C.R., Abbott, M., & Tapia, Y. (2006). Research on scaling up evidence-based instructional practice: Developing a sensitive measure of the rate of implementation. *Educational Technology Research and Development, 54*(5), 467-492. doi: 10.1007/s11423-006-0129-5

Campbell, J.I.D., & Thompson, V.A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods, 44*, 1255-1265. doi: 10.3758/s13428-012-0186-0

Cantrell, S.C., Almasi, J.F., Carter, J.C., & Rintamaa, M. (2013). Reading intervention in middle and high schools: Implementation fidelity, teacher efficacy, and student achievement. *Reading Psychology, 34*(1), 26-58.

Cao, J., Stokes, S.L., & Zhang, S. (2010). A Bayesian approach to ranking and rater evaluation: An application to grant reviews. *Journal of Educational and Behavioral Statistics, 35*, 194-214. doi: 10.3102/1076998609353116

Carlson, D., Borman, G.D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33*, 378-398. doi: 10.3102/0162373711412765

Caudle, K. (2010). Searching algorithm using Bayesian updates. *Journal of Computers in Mathematics and Science Teaching, 29*(1), 19-29.

Chaparro, E.A., Smolkowski, K., Baker, S.K., Hanson, N., & Ryan-Jackson, K. (2012). A model for system-wide collaboration to support integrated social behavior and literacy evidence-based practices. *Psychology in Schools, 49*, 465-482.

Choi, J., Kim, S., Chen, J., & Dannels, S. (2011). A comparison of Maximum Likelihood and Bayesian estimation for polychoric correlation using Monte Carlo simulation. *Journal of Educational and Behavioral Statistics, 36*, 523-549. doi: 10.3102/1076998610381398

Coburn, C.E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3-12.

Coffey, J.H., & Horner, R.H. (2012). The sustainability of Schoolwide Positive Behavior Interventions and Supports. *Exceptional Children, 78*, 407-422.

Cohen, R., Kincaid, D., & Childs, K.E. (2007). Measuring school-wide positive behavior support implementation: Development and validation of the benchmarks of quality. *Journal of Positive Behavior Interventions, 9*(4), 203-213. doi: 10.1177/10983007070090040301

Crawford, L., Carpenter, D.M., II, Wilson, M.T., Schmeister, M., & McDonald, M. (2012). Testing the relation between fidelity of implementation and student outcomes in math. *Assessment for Effective Intervention, 37*(4), 224-235.

Datnow, A., Park, V., & Wohlstetter, P. (2007). Achieving with data: How high-performing school systems use data to improve instruction for elementary students. Los Angeles: University of Southern California, Center of Educational Governance.

de Vries, R.M., & Morey, R.D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods, 18*, 165-185. doi: 10.1037/a0031037

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods, 18*, 186-219. doi: 10.1037/a0031609

Deublein, M., Schubert, M., Adey, B.T., Köhler, J., & Faber, M.H. (2013). Prediction of road accidents: A Bayesian hierarchical approach. *Accident Analysis and Prevention, 51*, 274-291. doi: 10.1016/j.aap.2012.11.019

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association, 100*(469).

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Enders, C.K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Ervin, R.A., Schaughency, E., Matthews, A., Goodman, S.D., & McGlinchey, M.T. (2007). Primary and secondary prevention of behavior difficulties: Developing a data-informed problem-solving model to guide decision making at a school-wide level. *Psychology in the Schools, 44*(1), 7-18. doi: 10.1002/pits.20201

Fixsen, D.L., Blase, K.A., Timbers, G.D., & Wolf, M.M. (2001). In search of program implementation: 792 replications of the Teaching Family Model. In G. A. Bernfeld, D. P. Farrington & A. W. Leschied (Eds.), *Offender rehabilitation in practice: Implementing and evaluating effective programs* (pp. 149-166). New York, NY, US: John Wiley & Sons Ltd.

Fixsen, D.L., Naoom, S.F., Blase, K.A., Friedman, R.M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Florida: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Flannery, K.B., Fenning, P., Kato, M.M., & McIntosh, K. (in press). Effects of School-Wide Positive Behavioral Interventions and Supports and fidelity of implementation on problem behavior in high schools. *School Psychology Quarterly*. doi: 10.1037/spq0000039

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional Item Response Theory model for Differential Item Functioning analysis on testlet-based items. *Applied Psychological Measurement, 35*(8), 604-622.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC press.

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

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*, 189-211. doi: 10.1080/19345747.2011.618213

Gelman, A., & Shalizi, C.R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*(1), 8-38. doi: 10.1111/j.2044-8317.2011.02037.x

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: Chapman & Hall/CRC.

Gill, J. (2009). *Bayesian methods: A social and behavioral sciences approach* (2nd ed.). Boca Raton, FL: CRC Press, Taylor & Francis Group.

Guskey, T.R. (1986). Staff development and the process of teacher change. *Educational Researcher, 15*(5), 5-12.

Han, S.S., & Weiss, B. (2005). Sustainability of teacher implementation of school-based mental health programs. *Journal of Abnormal Child Psychology, 33*(6), 665-679. doi: 10.1007/s10802-005-7646-2

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in education research. *Educational Researcher, 39*(2), 120-131. doi: 10.3102/0013189X10362578

Horner, R.H. (July, 2013). *Establishing, sustaining, and scaling effective practices*. Paper presented at the Office of Special Education Programs Project Directors Meeting, Washington, D.C.

Horner, R.H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A.W., & Esperanza, J. (2009). A randomized, wait-list controlled effectiveness trial assessing school-wide positive behavior support in elementary schools. *Journal of Positive Behavior Interventions, 11*(3), 133-144. doi: 10.1177/1098300709332067

Horner, R.H., Todd, A.W., Lewis-Palmer, T., Irvin, L.K., Sugai, G., & Boland, J.B. (2004). The School-Wide Evaluation Tool (SET): A research instrument for assessing school-wide positive behavior support. *Journal of Positive Behavior Interventions, 6*(1), 3-12. doi: 10.1177/10983007040060010201

Hoy, W., & Feldman, J. (1987). Organizational health: The concept and its measure. *Journal of Research and Development in Education, 20*, 30-38.

Hughes, R.I.G. (1997). Models and representation. *Philosophy of Science, 64*(4), 325-336.

Hulleman, C.S., & Cordray, D.S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*, 88-110. doi: 10.1080/19345740802539325

Hussong, A.M., Curran, P.J., & Bauer, D.J. (2013). Integrative Data Analysis in clinical psychology research. *Annual Review of Clinical Psychology, 9*(1), 61-89. doi: doi:10.1146/annurev-clinpsy-050212-185522

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics, 23*(1), 35-56.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82-100.

Kelly, D.L., Letson, D., Nelson, F., Nolan, D.S., & Solís, D. (2012). Evolution of subjective hurricane risk perceptions: A Bayesian approach. *Journal of Economic Behavior & Organization, 81*(2), 644-663. doi: 10.1016/j.jebo.2011.10.004

Kerr, K.A., Marsh, J.A., Ikemoto, G.S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education, 112*, 496-520.

Kincaid, D., Childs, K., Blase, K.A., & Wallace, F. (2007). Identifying barriers and facilitators in implementing schoolwide positive behavior support. *Journal of Positive Behavior Interventions, 9*, 174-184. doi: 10.1177/10983007070090030501

Kincaid, D., Childs, K., & George, H. (2005). School-wide benchmarks of quality. Tampa, FL: University of South Florida.

Kline, R.B. (2011). *Principals and practice of structural equation modeling* (3 ed.). New York: Guilford Press.

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods, 15*, 281-299. doi: 10.1037/a0020137

Kruschke, J.K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology, 142*, 573-603. doi: 10.1037/a0029146

Kuiper, R.M., Buskens, V., Raub, W., & Hoijtink, H. (2013). Combining statistical evidence from several studies: A method using Bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods & Research, 42*(1), 60-81. doi: 10.1177/0049124112464867

Larsen, R.J., & Marx, M.L. (2001). *Introduction to Mathematical Statistics and Its Applications* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Liu, C.C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52*, 362-375. doi: 10.1016/j.jmp.2008.03.002

Lohrmann, S., Forman, S., Martin, S., & Palmieri, M. (2008). Understanding school personnel's resistance to adopting schoolwide positive behavior support at a universal level of intervention. *Journal of Positive Behavior Interventions, 10*(4), 256-269. doi: 10.1177/1098300708318963

Luiselli, J.K., Putnam, R.F., Handler, M.W., & Feinberg, A.B. (2005). Whole-School Positive Behaviour Support: Effects on student discipline problems and academic performance. *Educational Psychology, 25*(2-3), 183-198. doi: 10.1080/0144341042000301265

Lunn, D.J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-327.

Masson, M.E.J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*, 679-690. doi: 10.3758/s13428-010-0049-5

Mathews, S., McIntosh, K., Frank, J.L., & May, S.L. (in press). Critical features predicting sustained implementation of School-Wide Positive Behavioral Interventions and Supports. *Journal of Positive Behavior Interventions*.

McIntosh, K., Doolittle, J.D., Vincent, C.G., Horner, R.H., & Ervin, R.A. (2009). School-wide universal behavior sustainability index: School teams. Vancouver, Canada: University of British Columbia.

McIntosh, K., Filter, K.J., Bennett, J.L., Ryan, C., & Sugai, G. (2010). Principles of sustainable prevention: Designing scale-up of school-wide positive behavior support to promote durable systems. *Psychology in the Schools, 47*(1), 5-21.

McIntosh, K., MacKay, L.D., Hume, A.E., Doolittle, J., Vincent, C.G., Horner, R.H., & Ervin, R.A. (2011). Development and initial validation of a measure to assess factors related to sustainability of school-wide positive behavior support. *Journal of Positive Behavior Interventions, 13*(4), 208-218. doi: 10.1177/1098300710385348

Mcintosh, K., Mercer, S.H., Hume, A.E., Frank, J.L., Turri, M.G., & Mathews, S. (2013). Factors related to sustained implementation of Schoolwide Positive Behavior Support. *Exceptional Children, 79*(3), 293-311.

McIntosh, K., Predy, L.K., Upreti, G., Hume, A.E., Turri, M.G., & Mathews, S. (in press). Perceptions of contextual features related to the implementation and sustainability of School-Wide Positive Behavior Support. *Journal of Positive Behavior Interventions*.

McIntosh, K., Sadler, C., & Brown, J.A. (2012). Kindergarten reading skill level and change as risk factors for chronic problem behavior. *Journal of Positive Behavior Interventions, 14*(1), 17-28. doi: 10.1177/1098300711403153

Mehta, P.D., & West, S.G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5*, 23-43. doi: 10.1037/1082-989X.5.1.23

Mercer, S.H., McIntosh, K., Strickland-Cohen, M.K., & Horner, R.H. (Manuscript submitted). Factorial invariance of a measure assessing sustainability of school-based universal behavior practices.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*(2), 177-195. doi: 10.1007/BF02293979

Morey, R.D., & Morey, C.C. (2011). WoMMBAT: A user interface for hierarchical Bayesian estimation of working memory capacity. *Behavior Research Methods, 43*, 1044-1065. doi: 10.3758/s13428-011-0114-8

Morey, R.D., Romeijn, J.W., & Rouder, J.N. (2013). The humble Bayesian: Model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology, 66*(1), 68-75.

Morey, R.D., & Rouder, J.N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*, 406-419. doi: 10.1037/a0024377

Muthén, B.O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313-335. doi: 10.1037/a0026802

Muthén, L.K., & Muthén, B.O. (1998-2012). *Mplus user's guide, seventh edition*. Los Angeles, CA: Muthén & Muthén.

Natarajan, R., & McCulloch, C.E. (1998). Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *Journal of Computational and Graphical Statistics, 7*(3), 267-277. doi: 10.1080/10618600.1998.10474776

National Center for Education Statistics. (2011). Common core of data: Public elementary/secondary school universe survey 2009-10 [Version 1a]. from http://nces.edu.gov/ccd/pucschuniv.asp

Newton, J.S., Algozzine, B., Algozzine, K., Horner, R.H., & Todd, A.W. (2011). Building local capacity for training and coaching data-based problem solving with positive behavior intervention and support teams. *Journal of Applied School Psychology, 27*, 228-245. doi: 10.1080/15377903.2011.590104

Newton, J.S., Horner, R.H., Algozzine, B., Todd, A.W., & Algozzine, K. (2012). A randomized wait-list controlled analysis of the implementation integrity of team-initiated problem solving processes. *Journal of School Psychology, 50*(4), 421-441.

O'Donnell, C.L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

O'Flaherty, B., & Komaki, J.L. (1992). Going beyond with Bayesian updating. *Journal of Applied Behavior Analysis, 25*(3), 585-597. doi: 10.1901/jaba.1992.25-585

Odom, S.L. (2009). The tie that binds evidence-based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education, 29*(1), 53-61.

Office of Student Achievement and School Accountability Programs. (2011). Improving basic programs operated by local educational agencies (Title I, Part A). Retrieved March 11, 2014, from http://www2.ed.gov/programs/titleiparta/index.html

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical model using Gibbs sampling.* Paper presented at the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

R Core Team. (2012). R: A language and environment for statistical computing. Austria: R Foundation for Statistical Comuting. Retrieved from http://www.R-project.org/

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). California: Sage.

Reinke, W.M., Herman, K.C., Stormont, M., Newcomer, L., & David, K. (2013). Illustrating the multiple facets and levels of fidelity of implementation to a teacher classroom management intervention. *Administration and Policy in Mental Health and Mental Health Services Research, 40*(6), 494-506. doi: 10.1007/s10488-013-0496-2

Rohrbach, L.A., Graham, J.W., & Hansen, W.B. (1993). Diffusion of a school-based substance abuse prevention program: Predictors of program implementation. *Preventive Medicine: An International Journal Devoted to Practice and Theory, 22*(2), 237-260. doi: 10.1006/pmed.1993.1020

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

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*, 229-237. doi: 10.3102/1076998609332756

Santangelo, T. (2009). Collaborative problem solving effectively implemented, but not sustained: A case for aligning the sun, the moon, and the stars. *Exceptional Children, 75*, 185-209.

School-wide PBIS: What is school-wide PBIS? (2013). Retrieved October 20, 2013, from http://www.pbis.org/school/

Schulte, A.C., Easton, J.E., & Parker, J. (2009). Advances in treatment integrity research: Multidisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review, 38*(4), 460-475.

Scott, T.M., & Barrett, S.B. (2004). Using staff and student time engaged in disciplinary procedures to evaluate the impact of school-wide PBS. *Journal of Positive Behavior Interventions, 6*(1), 21-27. doi: 10.1177/10983007040060010401

Shadish, W.R., Rindskopf, D.M., Hedges, L.V., & Sullivan, K.J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavior research methods, 45*, 813-821. doi: 10.3758/s13428-012-0282-1

Simonsen, B., Eber, L., Black, A.C., Sugai, G., Lewandowski, H., Sims, B., & Myers, D. (2012). Illinois statewide Positive Behavioral Interventions and Supports: Evolution and impact on student outcomes across years. *Journal of Positive Behavior Interventions, 14*(1), 5-16.

Sinharay, S., Dorans, N.J., Grant, M.C., & Blew, E.O. (2009). Using past data to enhance small sample DIF estimation: A Bayesian approach. *Journal of Educational and Behavioral Statistics, 34*(1), 74-96. doi: 10.3102/1076998607309021

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35*(2), 137-167. doi: 10.1207/S15327906MBR3502_1

Slavin, R.E., Holmes, G., Madden, N.A., Chamberlain, A., Cheung, A., & Borman, G. (2010). Effects of a data-driven district-level reform model: Center for Research and Reform in Education.

Soares, T.M., Gonçalves, F.B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics, 34*(3), 348-377. doi: 10.3102/1076998609332752

Sparks, G.M. (1988). Teachers' attitudes toward change and subsequent improvements in classroom teaching. *Journal of Educational Psychology, 80*(1), 111-117.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: B, 64*, 583-639.

Stone, C.A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*(1), 1-26.

Su, Y.-S., Gelman, A., Hill, J., & Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software, 45*(2), 1-31.

Sugai, G., Horner, R.H., & Lewis-Palmer, T. (2002). Team Implementation Checklist (version 2.2). Eugene, OR: Educational & Community Supports, University of Oregon.

Sugai, G., Horner, R.H., & Lewis-Palmer, T. (2009). Team Implementation Checklist (v. 3.0). Eugene, OR: Educational & Community Supports, University of Oregon.

Sugai, G., Horner, R.H., Lewis-Palmer, T., & Rossetto Dickey, C. (2011). Team Implementation Checklist (version 3.1). Educational and Community Supports, University of Oregon.

Sugai, G., Lewis-Palmer, T.L., Todd, A.W., & Horner, R.H. (2001). School-wide evaluation tool. Eugene, OR: Educational and Community Supports.

Sugai, G., Todd, A.W., & Horner, R.H. (2001). Team implementation checklist. Eugene, OR: OSEP Center for Positive Behavioral Supports.

Swaminathan, H., & Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*(3), 175-191. doi: 10.2307/1164643

Swaminathan, H., & Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*, 349-364. doi: 10.1007/BF02294110

Swaminathan, H., & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589-601. doi: 10.1007/BF02295598

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention fidelity in special and general education research journals. *Journal of Special Education, 47*(1), 3-13.

The Stan Development Team. (2013). Stan modeling language user's guide and reference manual. Retrieved July 15, 2013, from https://code.google.com/p/stan/wiki/RStanGettingStarted#RStan_Getting_Started

Tobin, T.J. (2006). *Implementing positive behavior support in regular and alternative high schools: Use of the Team Implementation Checklist*. Paper presented at the Association for Positive Behavior Support's Third International Conference on Positive Behavior Support, Reno, N.V. http://pages.uoregon.edu/ttobin/alt_tic.pdf

Tobin, T.J., Vincent, C.G., Horner, R.H., Rossetto Dickey, C., & May, S.A. (2012). Fidelity measures to improve implementation of behavioural support. *International Journal of Positive Behavioural Support, 2*(2), 12-19.

Todd, A.W., Horner, R.H., Newton, J.S., Algozzine, R.F., Algozzine, K.M., & Frank, J.L. (2011). Effects of team-initiated problem solving on decision making by schoolwide behavior support teams. *Journal of Applied School Psychology, 27*(1), 42-59. doi: 10.1080/15377903.2011.540510

Vanpaemel, W. (2009). BayesGCM: Software for Bayesian inference with the generalized context model. *Behavior Research Methods, 41*, 1111-1120. doi: 10.3758/BRM.41.4.1111

Vaughn, S., Klingner, J., & Hughes, M. (2000). Sustainability of research-based practices. *Exceptional Children, 66*, 163-171.

Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics, 35*(1), 5-25. doi: 10.3102/1076998609355124

Wang, Y., & Daniels, M.J. (2013). Bayesian modeling of the dependence in longitudinal data via partial autocorrelations and marginal variances. *Journal of Multivariate Analysis, 116*, 130-140. doi: 10.1016/j.jmva.2012.11.010

Wetzels, R., Lee, M.D., & Wagenmakers, E.-J. (2010). Bayesian inference using WBDev: A tutorial for social scientists. *Behavior Research Methods, 42*, 884-897. doi: 10.3758/BRM.42.3.884

Witt, J.C., Martens, B.K., & Elliott, S.N. (1984). Factors affecting teachers' judgments of the acceptability of behavioral interventions: Time involvement, behavior problem severity, and type of intervention. *Behavior Therapy, 15*(2), 204-209. doi: 10.1016/S0005-7894(84)80022-2

Yu, R., & Abdel-Aty, M. (2013). Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions. *Accident Analysis and Prevention, 56*, 51-58. doi: 10.1016/j.aap.2013.03.023

Yuan, Y., & MacKinnon, D.P. (2009). Bayesian mediation analysis. *Psychological Methods, 14*, 301-322. doi: 10.1037/a0016972

Zeng, L. (1997). Implementation of marginal Bayesian estimation with four-parameter beta prior distributions. *Applied Psychological Measurement, 21*(2), 143-156. doi: 10.1177/01466216970212004

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. *American Journal of Evaluation, 30*(1), 44-61. doi: 10.1177/1098214008329523

Zvoch, K. (2012). How does fidelity of implementation matter? Using multilevel models to detect relationships between participant outcomes and the delivery and receipt of treatment. *American Journal of Evaluation, 33*(4), 547-565. doi: 10.1177/1098214012452715

Zvoch, K., Letourneau, L.E., & Parker, R.P. (2007). A multilevel multisite outcomes-by-implementation evaluation of an early childhood literacy model. *American Journal of Evaluation, 28*(2), 132-150. doi: 10.1177/1098214007301138

Zvoch, K., & Stevens, J.J. (2006). Longitudinal effects of school context and practice on middle school mathematics achievement. *The Journal of Educational Research, 99*(6), 347-356. doi: 10.3200/JOER.99.6.347-357

Zwick, R., Ye, L., & Isham, S. (2012). Improving Mantel–Haenszel DIF estimation through Bayesian updating. *Journal of Educational and Behavioral Statistics, 37*, 601-629. doi: 10.3102/1076998611431085