Essays on Treatment Effect Heterogeneity in Education Policy Interventions

by

Joon-Ho Lee

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sophia Rabe-Hesketh, Chair
Professor Bruce Fuller
Associate Professor Avi Feller

Spring 2020

Essays on Treatment Effect Heterogeneity in Education Policy Interventions

Abstract

Essays on Treatment Effect Heterogeneity in Education Policy Interventions

by

Joon-Ho Lee

Doctor of Philosophy in Education

University of California, Berkeley

Professor Sophia Rabe-Hesketh, Chair

The key focus of this dissertation is on how to understand and measure treatment effect heterogeneity in experimental or quasi-experimental evaluations of educational policy interventions. When testing the impact of an intervention, it can be important to know not just the overall or average effect of the intervention on key outcomes but also how the effect varies across subgroups of study participants, as defined by several dimensions including their pre-treatment characteristics, site-level contexts, and the distribution of an outcome measure. Heterogeneity or variation in effects has critical implications for understanding how interventions work and which aspects of an intervention's implementation are most closely associated with its effectiveness. This dissertation examines both methodological and substantive questions that pertain to such heterogeneity.

In Chapter 1, I examine Bayesian hierarchical models for multi-site trials that allow estimation of site-specific treatment effects and their distribution. Modeling site-specific effects using observed data is a critical component in understanding the results of multisite trials. A standard approach leveraging Bayesian methods is to rely on Gaussian distributional assumptions and to use the posterior means (PM) of the random effects. The standard approach can be misleading, however, in the estimation of individual site-specific effects and their empirical distribution and ranks. In this chapter, I review the following two strategies developed to improve inferences regarding site-specific effects: (a) relaxing the normality assumption by flexible modeling of the random-effects distribution using Dirichlet process mixture (DPM) models, and (b) replacing the choice of PM as the summary of the posterior by alternative estimators, such as the constrained Bayes (CB) or the triple-goal (GR) estimators. I then examine when and to what extent the two strategies and combinations thereof work or fail under varying conditions.

In Chapter 2, I study methodological issues arise in the practice where Bayesian quantile regression (BQR) models are applied. The BQR models allow us to study treatment effect heterogeneity across the distribution of an outcome measure such as a student achievement

test score. In BQR, the most commonly applied likelihood is the asymmetric Laplace (AL) likelihood because it is computationally convenient for Markov chain Monte Carlo algorithms. For easier computation, the scale parameter of the AL distribution is often fixed at a pre-estimated value or an arbitrary constant. This paper demonstrates that posterior inference in BQR with an AL likelihood is highly sensitive to the choice of the fixed scale parameter. Based on sensitivity analyses using Monte Carlo simulations and a real data example, I make two claims. First, not only the variance directly obtained from the posterior distribution, but also the adjusted posterior variance proposed by Yang et al. (2015), is highly sensitive to the value of the scale parameter. Second, in finite samples, both conventional and Bayesian point estimators can be biased at extreme quantiles. Researchers need to be aware of the possibility of low coverage probabilities at extreme quantiles mainly caused by biased point estimates.

In Chapter 3, I examine the use of the grouped/multilevel instrumental variable (IV) quantile regression approach, a quantile extension of Hausman and Taylor (1981). The common approach of estimating the shift of group-level (level-2) averages of individual-level (level-1) outcomes may mask important but more subtle effects on the outcome distribution. For example, a school-level intervention may have little effect on school-level average test score but may cause a substantial shift in the lower quantiles of the within-school test score distributions if the intervention is particularly beneficial for low-performing students. As one real-world empirical example, I used the grouped/multilevel IV quantile approach to estimate the effects of district-level increases in per-pupil spending on quantiles of the within-district distribution of school quality measures. I show how new dollars flowing to districts did affect varying mixes of teachers and organizational practices inside schools, but in ways that mitigated against narrowing disparities. Better funded high schools reduced access to college-prep courses relative to electives, and novice teachers were often assigned to courses serving English learners, inequities that widened in high-poverty schools.

To my late grandmother, Bong-Hee Lee (1944-2005)

Contents

Contents **ii**

List of Figures **iv**

List of Tables **vi**

1 Improving the Estimation of Site-Specific Effects and their Distribution in Multisite Trials **1**

1.1 Introduction 1

1.2 Standard approaches to modeling site-specific effects 3

1.3 Improving inferences for site-specific effects 6

1.4 A simulation study 13

1.5 Conclusion 31

2 Sensitivity of Bayesian Quantile Regression to the Scale Parameter of the Asymmetric Laplace Likelihood 34

2.1 Introduction 34

2.2 Bayesian quantile regression with asymmetric Laplace likelihood 36

2.3 Sensitivity of posterior inference to the scale parameter 39

2.4 Simulation studies 42

2.5 Real data analysis 54

2.6 Conclusion 58

3 Does finance Reform Move Teachers and School Organizations? Californias $23 Billion Equity Initiative 66

3.1 Introduction 66

3.2 Does finance reform alter school organizations? 68

3.3 Progressive finance reform in California 71

3.4 Analytic strategy 74

3.5 Method 77

3.6 Findings 89

3.7 Discussion and policy implications 100

Bibliography 103

List of Figures

1.1 The derivation of the informative prior for the precision parameter α by approximating a Gamma(a, b) to a $\chi^2(5)$ 13
1.2 The true population distribution G: Gaussian, mixture of two Gaussian, and asymmetric Laplace distributions 15
1.3 The distributions of the expected number of clusters K for the DP-diffuse and DP-inform models ($N = 50$) 17
1.4 Estimated empirical distribution function (EDF) when true population distribution of G is a mixture of two Gaussians 23
1.5 The effect of the number of sites (N) and the heterogeneity of $\hat{se}_j^2$ (R) on the two loss estimates: (a) the MSEL of τ_j estimated using PM and (b) the ISEL of the EDF estimated using GR 24
1.6 The effect of using CB or GR on the two loss estimates: (a) the MSEL of τ_j and (b) the ISEL of the EDF of τ_js 25
1.7 The effect of using DP priors on the two loss estimates: (a) the MSEL of τ_j estimated using PM and (b) the ISEL of the EDF estimated using GR 26
1.8 The effect of the number of sites (N) on the biases of the 50th and 90th percentiles of G 29
1.9 The effect of using the DP priors on the biases of the 50th and 90th percentiles of G 31

2.1 Standard ($\mu = 0, \sigma = 1$) asymmetric Laplace densities at $\tau = 0.1, 0.5, 0.9$ 37
2.2 Kernel density plot of 1,000 draws from the ALD and a scale mixture of normals with $\mu = 0$, $\sigma = 1$, and $\tau = 0.9$ 39
2.3 Sensitivity of bias to the fixed scale parameter σ in Case 1 with $n = 500$ 48
2.4 Sensitivity of the biases of $\hat{\alpha}(\tau)$ to the fixed scale parameter σ in Case 1 with $n = 200$ and $n = 500$ 48
2.5 Sensitivity of bias to the fixed scale parameter σ in Case 4 with $n = 500$ 49
2.6 Sensitivity of the biases of $\hat{\alpha}(\tau)$ to the fixed scale parameter σ in Case 4 with $n = 200$ and $n = 500$ 49
2.7 Sensitivity of the relative error of the model-based SE to the fixed scale parameter σ in Case 1 with $n = 500$ 51

2.8 Sensitivity of the relative error of the model-based SE to the fixed scale parameter σ in Case 4 with $n = 500$ 52
2.9 Sensitivity of the coverage of 95% confidence or credible intervals to the fixed scale parameter σ in Case 4 with $n = 500$ 53
2.10 Sensitivity of the estimates of $\alpha(\tau)$ and $\beta(\tau)$ to the fixed scale parameter σ . . . 56
2.11 Sensitivity of bias to the fixed scale parameter σ in Case 2 with $n = 500$ 59
2.12 Sensitivity of bias to the fixed scale parameter σ in Case 3 with $n = 500$ 60
2.13 Sensitivity of the relative error of the model-based SE to the fixed scale parameter σ in Case 2 with $n = 500$ 61
2.14 Sensitivity of the relative error of the model-based SE to the fixed scale parameter σ in Case 3 with $n = 500$ 62
2.15 Sensitivity of the coverage of 95% confidence or credible intervals to the fixed scale parameter σ in Case 1 with $n = 500$ 63
2.16 Sensitivity of the coverage of 95% confidence or credible intervals to the fixed scale parameter σ in Case 2 with $n = 500$ 64
2.17 Sensitivity of the coverage of 95% confidence or credible intervals to the fixed scale parameter σ in Case 4 with $n = 500$ 65

3.1 Change in high school mathematics Smarter Balanced state assessment (SBA) results by student subgroup 73
3.2 Three stages (or levels of social organization) in which inputs are distributed among districts, schools, and students 75
3.3 Observed and predicted (counterfactual) levels of California state revenues per pupil, 2003-2016. Note: Dollars are deflated into constant 2016 dollars 85
3.4 Time trends in selected teacher characteristics for low- and high-poverty elementary schools (Q1 and Q5, respectively), 2003-2017 90
3.5 Time trends for selected organizational features and curricular structure for low- and high-poverty elementary and high schools (Q1 and Q5, respectively), 2003-2017 91
3.6 Time trends in access to non-tenured teachers and college-prep (A-G) classes by English learners, 2003-2017 93
3.7 Event study estimates of Local Control Funding effects on school-level distributions of teacher characteristics (median, 0.5 quantile) 96
3.8 Event study estimates of Local Control Funding effects on school-level distributions of shares of white teachers and masters degree holders (0.2 and 0.8 quantiles) 97
3.9 Event study estimates of Local Control Funding effects on distributions of school-level organizational practices and curricular 98

List of Tables

1.1 Meta-model regression results (Data-generating model: a mixture of two Gaussians) 20
1.2 Meta-model regression results (Data-generating model: asymmetric Laplace distribution) 21
1.3 Meta-model regression results for mean biases in the 50th and 90th percentiles of G 28

2.1 The empirical estimates and their Monte Carlo errors of performance evaluators 44
2.2 Bias, relative error, and coverage probability in case 1 and 4 with $n = 500$ when the scale parameter is fixed at $\hat{\sigma}_{\tau=0.5}$ 46
2.3 The point estimates and model-based standard error estimates (in parentheses) of $\alpha(\tau)$ and $\beta(\tau)$ at $\tau = 0.1, 0.5, 0.9$ as σ varies in Bayesian quantile regression with AL likelihood 57

3.1 Descriptive statistics for teacher characteristics, organizational features, and curricular structure by low- and high-poverty schools (Q1 and Q5), pooled data, 2003-2016 79
3.2 Access of English learners to experienced teachers and college-prep (A-G) classes, pooled data, 2012-2017 81
3.3 Summary of the estimated difference-in-difference estimates of the effects of LCF-induced spending increases on school-level teacher characteristics, organizational features, and curricular structure 94

Chapter 1

Improving the Estimation of Site-Specific Effects and their Distribution in Multisite Trials

1.1 Introduction

Multisite trials, which arise when individuals are randomly assigned to experimental conditions within each site, are prevalent in education research. For example, multisite trials take up more than 66% of the 175 randomized controlled trials (RCTs) funded by the Institute of Education Sciences in the past 26 years (Spybrook, 2013; Raudenbush & Bloom, 2015). Designing and analyzing multisite trials is gaining increased interest because multisite trials can provide an opportunity to address a series of important questions about the effectiveness of educational interventions.

A multisite trial can be viewed as *a fleet of randomized experiments* or *a planned meta-analysis* (Bloom et al., 2017). Thus, it shares common inferential goals with a meta-analysis such as (a) estimating overall mean impact of treatment or (b) quantifying treatment effect heterogeneity across study sites. Estimating mean impact is of concern to aggregate evidence across multiple contexts to provide a reasonable basis for general policy recommendations (Meager, 2019). Knowing about the extent to which treatment effects vary offers an important ground for understanding how an intervention might differentially operate in different contexts (Miratrix et al., 2016).

Many inferential goals in multisite trials, however, require studying individual site-specific treatment effects. School or teacher effectiveness studies, for example, directly aim to estimate the individual school- or teacher-specific effect parameters (Raudenbush & Willms, 1995). There also has been continuous interests in producing rankings or *league tables* based on the estimated site-specific effects for schools or other service providers (Goldstein & Spiegelhalter, 1996; Lockwood et al., 2002; Normand & Shahian, 2007). Similar performance evaluation goals based on site-specific effects include identifying hot spots (Wright et al., 2003) or estimating the proportion of sites with an effect larger or smaller than some threshold (Conlon & Louis, 1999; Miratrix et al., 2016).

If the individual site-specific effect parameters could be observed, we could easily construct the true cross-site effect distribution and generate the target quantities according to it. As we cannot observe the true value of the site-specific effect, however, a key task in this setting is to model the distribution of the individual site-level effects using observed data. A standard modeling approach is to rely on parametric distributional assumptions in random-effect multilevel models and to use an empirical Bayes (EB) prediction of the random effects (e.g., Raudenbush & Bloom, 2015). When a fully Bayesian approach is adopted, as in this chapter, empirical Bayes, is replaced by the posterior means (PM) of the random effects over the joint posterior distribution of the random effects and model parameters. A convenient modeling assumption is that the distribution is Gaussian, but this could be problematic when the true distribution is multi-modal or long-tailed, for instance. McCulloch and Neuhaus (2011) pointed out that the shape of the distribution of the EB predictions is likely to reflect the assumed Gaussian distribution, not the true underlying distribution of the random effects.

A popular response to this threat has been to adopt flexible distributional assumptions for the random effects. Flexible alternatives include (a) continuous parametric non-Gaussian distributions such as Student's t (Pinheiro et al., 2001) or the skewed parametric family (Liu & Dey, 2008), (b) arbitrary discrete distributions through nonparametric maximum likelihood (NPML, Rabe-Hesketh et al., 2003) or smoothing by roughening (Shen & Louis, 1999), and (c) mixture distributions such as finite mixtures of Gaussians (Verbeke & Lesaffre, 1996), penalized Gaussian mixtures (Ghidey et al., 2004), or Dirichlet process mixtures (Paddock et al., 2006; Antonelli et al., 2016). These studies advocate the use of flexible distributional assumptions since they protect against model misspecification with little loss in efficiency.

Instead of relaxing the usual normality assumption by flexible modeling of the random-effects distribution, alternative approaches replace the choice of PM (or EB) as the summary of the posterior by alternative estimators, such as the *constrained Bayes* (Louis, 1984) or the *triple-goal* (Shen & Louis, 1998) estimators. These estimators have been developed to correct the underdispersion of the distribution of posterior mean estimates induced by shrinkage. Rather than focusing on the robustness of the model specification, these approaches modify the loss function being minimized by the estimator, targeting the loss function toward inferential goals. For example, the triple-goal estimators aim to optimize the estimation of the empirical distribution and ranks of the site-specific parameters with a little trade-off in the optimality of individual parameter estimates. These strategies have been considered less than flexible modeling of the random-effects distribution and have rarely been used jointly with flexible modeling (e.g., Antonelli et al., 2016).

Thus, the primary focus of this study is to investigate when and to what extent the two diverging strategies, flexible modeling of the prior distribution and minimizing alternative loss functions, and combinations thereof work or fail under varying conditions. Paddock et al. (2006) provides pioneering work on a similar issue, but they focus mainly on the performance of flexible Dirichlet process mixture (DPM) compared with the normality assumption combined when the triple-goal estimator is used. That is, the costs and benefits of using varied combinations of the two strategies with respect to different inferential goals were not considered. When the inferential goal is to recover the shape of the true distribution of site-specific effects, for instance, using a Gaussian model with the triple-goal estimator may be more effective than employing a flexible DPM model with the PM which induces underdispersion of the distribution due to shrinkage.

We are particularly interested in the *low-data environment* where the number of sites or the number of individuals within sites is hardly large. In the context of multisite trials, sites are generally small to moderate in number, reaching only to the hundreds (Miratrix et al., 2016). For example, the number of sites was equal to 5 in the Moving to Opportunity experiment (Katz et al., 2000), 19 in the Early College High Schools study in North Carolina (Edmunds et al., 2015), 65 in the National Study of Learning Mindsets study (Yeager et al., 2019), and 350 in the national Head Start Impact Study (Bloom & Weiland, 2015). Furthermore, even large-scale multisite trials are likely to include a high proportion of sites with small sizes which leads to large sampling variation. Paddock et al. (2006)'s well-designed simulation study did not consider the number of sites as a factor and considered only moderate to high levels of sampling error which correspond to the conventional pooling factor metrics (Gelman & Hill, 2007) larger than 0.5. Here we vary the number of sites, consider a larger range of sampling errors, and directly compare the influences of the two strategies for improving inferences regarding site-specific effects.

This chapter is organized as follows. We begin by presenting standard approaches to modeling site-specific effects, namely the Rubin (1981) model and the posterior mean estimator. We then discuss inferential goals and threats to inferences for site-specific effects and provide a detailed description of the two strategies to improve inferences for a distribution of site-specific effects. Next, we provide the design and results of our simulation study. Finally, we summarize simulation results and discuss their implications.

1.2 Standard approaches to modeling site-specific effects

1.2.1 Basic setup: the Rubin (1981) model

Multisite trials generate multilevel or clustered data because individuals are randomly assigned to a treatment or control group within each site. While a broad set of generalized linear mixed models is available to analyze such multilevel data, this paper focuses on the Rubin (1981) model for parallel randomized experiments, also known as a random-effects (or empirical Bayes) meta-analysis (DerSimonion & Laird, 1986; Raudenbush & Bryk, 1985). Suppose a multisite trial consists of N sites indexed by $j = 1, \dots, N$ in which the same treatments are performed. Since site-specific *true* effects, τ_j's, are unobservable, researchers only have access to the observed or estimated effects $\hat{\tau}_j$ from each of the N sites with their corresponding squared standard errors $\widehat{se}_j^2$. The $\hat{\tau}_j$ and $\widehat{se}_j^2$ are obtained by maximum likelihood (ML) estimation using only the data from site j. The first stage of Rubin's (1981) hierarchical model describes the relationship between the observed data $\hat{\tau}_j$ and the latent parameter τ_j:

$$\hat{\tau}_j | \tau_j, \widehat{se}_j^2 \sim N\left(\tau_j, \widehat{se}_j^2\right) \quad j = 1, \dots, N. \tag{1}$$

The second stage of the Rubin model assumes that τ_j are independent and identically distributed ($i.i.d.$) with a certain prior distribution G,

$$\tau_j | \tau, \sigma^2 \sim G \equiv N(\tau, \sigma^2) \quad j = 1, \dots, N. \tag{2}$$

The prior distribution G is unknown in general, but the Rubin (1981) model specifies G as a Gaussian distribution with two hyperparameters: τ, the mean treatment effect, and σ^2, the variance in true site-specific effects τ_j's across sites, both defined at the population level.

The key insight of the Rubin model is that the observed variation in estimated site-specific effects, $\mathrm{var}(\hat{\tau}_j)$, reflects two sources of variation: (1) genuine heterogeneity in true effects τ_j between sites (σ^2), and (2) the sampling variation of each $\hat{\tau}_j$ around its τ_j within sites ($\widehat{se}_j^2$). When site sample sizes are small, the ML estimates of site-specific effects $\hat{\tau}_j$ can have large sampling error variances $\widehat{se}_j^2$. Thus, the $\hat{\tau}_j$ will have an empirical distribution function (EDF) that is very different from that of the true τ_j due to overdispersion of the EDF of $\hat{\tau}_j$. Furthermore, the rank order of effects for different sites can be misrepresented because sites with the smallest samples tend to have the most extreme estimates due to large $\widehat{se}_j^2$ (Raudenbush & Bloom, 2015). Hence, it is necessary to remove the influence of sampling error within sites to uncover the true heterogeneity in treatment effects across the population of sites. The Rubin (1981) model's hierarchical framework is designed to separate the genuine heterogeneity σ^2 from the sampling variation $\widehat{se}_j^2$ (Meager, 2019).

1.2.2 Site-specific parameter estimation

Maximum likelihood estimation (MLE), restricted maximum likelihood (REML) or Bayesian approaches are typically used to estimate the parameters. Bayesian approaches yield posterior mean estimates of τ, σ^2 and the τ_j, whereas MLE and REML yield estimates of τ, σ^2 only, and these parameters are treated as known when obtaining conditional posterior means of τ_j, also known as empirical Bayes (EB) estimates. In multisite studies and meta-analyses, the primary parameters of interest are typically τ and σ^2. Since the estimation of τ and σ^2 using Gaussian hierarchical models is found to be robust to misspecification of the prior distribution G (McCulloch & Neuhaus, 2011), the Rubin model can deliver reliable inference for these parameters.

The focus of this chapter is on inferences for τ_j, however, which can be sensitive to misspecification of G. Here we provide some details on the conditional posterior distribution of τ_j, given the hyperparameters τ and σ^2. When MLE or REML estimates for these hyperparameters are plugged in, the mean of the conditional posterior distribution is an empirical Bayes estimate.

Under the model's normality assumptions, the conditional posterior distribution of $\tau_j$is normal (Gelman et al., 2013)

$$\tau_j | \tau, \sigma^2, \hat{\tau}_j \sim N(\tau_j^*, V_j) \quad j = 1, \ldots, N,$$

where

$$\tau_j^* = \frac{\frac{1}{\sigma^2} \cdot \tau + \frac{1}{\widehat{se}_j^2} \cdot \hat{\tau}_j}{\frac{1}{\sigma^2} + \frac{1}{\widehat{se}_j^2}}, \quad V_j = \frac{1}{\sigma^2} + \frac{1}{\widehat{se}_j^2} \ . \tag{3}$$

The inverse of the σ^2 and $\widehat{se}_j^2$, the so-called *precisions*, hence serve a critical role in obtaining the conditional posterior mean τ_j^* and variance V_j. The conditional posterior mean effect is a weighted average of the prior mean effect τ and the observed effect for the site, $\hat{\tau}_j$, with weights given by the precisions.

The posterior mean effect τ_j^* can be rewritten as the observed mean effect $\hat{\tau}_j$ shrunk toward the the prior mean effect τ:

$$\tau_j^* = \tau + (\hat{\tau}_j - \tau) \cdot \frac{\sigma^2}{\sigma^2 + \widehat{se}_j^2} \ . \tag{4}$$

The weight, $\sigma^2/(\sigma^2 + \widehat{se}_j^2)$, can be interpreted as the *reliability* of the ML estimator of $\hat{\tau}_j$, defined as the proportion of the variance of the ML estimator that is due to the genuine underlying heterogeneity across sites (Rabe-Hesketh & Skrondal, 2012). If $\widehat{se}_j^2 = 0$, the ML estimator of $\hat{\tau}_j$ is perfectly precise or reliable, and thus the posterior mean and the ML estimator are identical ($\tau_j^* = \hat{\tau}_j$). A large $\widehat{se}_j^2$ indicates relatively less informative data about the τ_j than the prior distribution for τ, which results in the posterior mean effect τ_j^* shrunken more toward the prior mean effect τ. We use the weight, also known as the *shrinkage factor,* which ranges from 0 to 1, to compare the magnitude of σ^2 to that of $\widehat{se}_j^2$. If the weight is smaller than 0.5, it indicates that $\widehat{se}_j^2$ is smaller than σ^2, suggesting larger shrinkage toward the prior mean effect τ.

1.3 Improving inferences for site-specific effects

1.3.1 Inferential goals and threats to inferences for τ_j

If site-specific effects τ_j are the central parameters of interest, there can be three different inferential goals (Shen & Louis, 1998): (1) estimating the individual site-specific effect parameters, τ_j, (2) ranking the sites based on τ_j, and (3) estimating the empirical distribution function (EDF) of the τ_j's. For the first goal, we will explain below that it makes sense to use the posterior mean of τ_j as its estimator, but for the other goals, other summaries of the posterior distribution are preferable. Shen and Louis (1998) point out that the loss function that are minimized by the estimators should be targeted towards the inferential goal. For the first goal, the estimator with the least mean squared error loss (MSEL) is preferred:

$$\text{MSEL} = \frac{1}{N} \cdot \sum_{j=1}^{N} \left(a_j - \tau_j\right)^2 \ , \tag{5}$$

where a_j is the estimate of τ_j generated by a candidate estimator. The *posterior mean* (PM) of τ_j minimizes the MSEL. When the hyperparameters are treated as known, the conditional PM, such as τ_j^* in equation (3) and (4), is optimal with respect to the MSEL. That is, the conditional MSEL is minimized when $a_j = \tau_j^*$.

For the second inferential goal, we aim to identify an estimator for the vector of ranks of τ_j that minimizes the mean squared error loss of the ranks (MSELR),

$$\text{MSELR} = \frac{1}{N} \cdot \sum_{j=1}^{N} \left(\mathbf{T}_\mathrm{j} - \mathbf{R}_\mathrm{j}\right)^2 \ , \tag{6}$$

where $\mathbf{R}_\mathrm{j} = \sum_{j=1}^{N} I(\tau_i \geq \tau_j)$ is the true rank of τ_j with the indicator function $I(\cdot)$, and $\mathbf{T}_\mathrm{j}$ is the candidate vector of estimated ranks. As Goldstein and Spiegelhalter (1996) have shown, ranks based on the PMs can be suboptimal in general.

The Rubin model, combined with PM estimation of the site-specific effects, can perform poorly particularly for the third inferential goal. The third goal, estimating the empirical distribution function (EDF) of the τ_j's, is one of the Bayes *deconvolution* problems: using the observed sample $\hat{\tau}_j$ to recover an unknown prior density G (Laird, 1978; Stefanski & Carroll, 1990; Efron, 2016). The prior density G is not the same as the EDF of τ_j, but an estimator of the EDF of τ_j can be viewed as an estimator of G. If the PM estimator of the Gaussian hierarchical model is used for this goal, there can be multiple threats to the valid estimation of the prior distribution G.

First, even when the prior distribution G is correctly specified, it is well-known that the EDF of the PM effect estimates τ_j^* is under-dispersed relative to the EDF of τ_j because of the shrinkage toward the prior mean effect τ while the EDF of the observed ML effect estimates $\hat{\tau}_j$ is over-dispersed due to the presence of sampling errors (Mislevy et al., 1992). To resolve this issue, Shen and Louis (1998) suggest using the integrated squared error loss (ISEL) function to identify the optimal estimate of the EDF. Suppose the true EDF is $G_N(t) = N^{-1} \cdot \sum I_{\{\tau_j \leq t\}}$ where $-\infty < t < \infty$. Then the ISEL measures the discrepancy between the $G_N(t)$ and $A(t)$, a candidate estimator of $G_N(t)$:

$$\mathrm{ISEL(A, G_N)} = \int \{A(t) - G_N(t)\}^2 \mathrm{d}t \ . \tag{7}$$

Second, the shape of the estimated EDF of τ_j's can be sensitive to the assumed form of the prior distribution G. McCulloch and Neuhaus (2011) have shown that most aspects of statistical inference are highly robust to the misspecified Gaussian assumption for the random effects. According to their simulation study, however, the shape of the estimated random effects distribution was one prominent exception to the robustness. If the true prior distribution for G is not Gaussian, the adoption of a Gaussian prior for G leads to a misspecified likelihood function, which results in nonresponsiveness to skewness, long-tail, multimodality and other complexities in the estimation of the EDF of the τ_j's. Since there hardly exist any substantive reasons to believe that the true distribution of site-specific effects follows a Gaussian distribution, the Rubin model assuming the Gaussian prior for G can be unreliable for the third inferential goal, particularly when estimating thresholds or tails of the underlying prior distribution for the τ_j's.

1.3.2 Strategies to improve inferences for a distribution of τ_j

There have been two strategies to respond to threats explained in the previous section, one regarding posterior sample summarization and the other regarding specification for the prior distribution G. The first is to use posterior summary methods that are directly targeted to an inferential goal via choice of the appropriate loss function. The posterior mean is one kind of posterior summary estimators which aims to minimize the MSEL of the individual τ_j. To minimize the ISEL of the EDF of τ_j's, Raudenbush and Bloom (2015) recommend using *constrained Bayes* estimator (Louis, 1984; Ghosh, 1992) which rescales the posterior means to have variance equal to the estimated marginal variance of the τ_j's. The triple-goal estimator developed by Shen and Louis (1998) also directly addresses the threats by attempting to balance trade-offs between the losses for the three inferential goals. The section 1.3.3 provides a detailed explanation of these estimators.

The second strategy is to adopt flexible semiparametric or nonparametric specifications for the prior distribution G to protect against model misspecification (Paddock et al., 2006). To relax the Gaussian assumption for G, we can hypothesize a less restrictive prior distribution for G that specifies the space of distributions that G can take on and specifies a prior for the distributions in the selected space (Antonelli et al., 2016). The Dirichlet process (DP) prior is one of the most commonly used nonparametric specifications among numerous such priors proposed in the Bayesian literature (Lockwood et al., 2018). We explain the use of the DP prior in the current setting in section 1.3.4.

These two strategies have been rarely used jointly in practice with some notable exceptions (e.g., Paddock et al., 2006; Lockwood et al., 2018). In addition, the costs and benefits of the two strategies have not been compared directly much in detail under varying conditions in the previous simulation studies. The benefits of the strategies on recovering an unknown prior density G may differ depending upon the number of sites, the reliability of the ML estimator of $\hat{\tau}_j$, the heterogeneity of the $\widehat{se}_j^2$, and the shape of true population distribution of G. In particular, we are unaware of any previous studies that investigate the effect of having a small to moderate number of sites which is common in the context of meta-analyses and multisite trials. Models with DP prior, for example, may requires sample sizes that are quite large to decently recover G or may be highly sensitive to the specification of hyperpriors.

1.3.3 Posterior summary methods: constrained Bayes and triple-goal estimators

In this section, we consider two posterior summary methods that have been developed to respond to the threat posed by under-dispersion of PMs: the constrained Bayes estimator (Louis, 1984; Ghosh, 1992) and triple-goal goal estimator (Shen & Louis, 1998). Our target of interest is G_N, the EDF of τ_j's. Shen and Louis (1998) showed that the optimal EDF estimator that that minimizes the ISEL in equation (7) is

$$\bar{G}_N(t) = \mathrm{E}[\mathrm{G_N}(t) \mid \hat{\boldsymbol{\tau}}] = \frac{1}{N} \cdot \sum \Pr\left(\tau_j \le t \middle| \hat{\tau}_j\right), \tag{8}$$

where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \cdots, \hat{\tau}_N)$. Let the posterior mean of τ_j, $\mathrm{E}(\tau_j|\hat{\tau}_j)$, be η_j and the posterior variance of τ_j, $\mathrm{Var}(\tau_j|\hat{\tau}_j)$, be λ_j. Then, the marginal mean of $\bar{G}_N$ and the finite sample version of the marginal variance of $\bar{G}_N$ can be defined as follows (Shen & Louis, 1998):

$$\mathrm{E}[\bar{G}_N] = \int t \mathrm{d}\, \bar{G}_N(t) = \frac{\sum \eta_j}{N} = \bar{\eta}\,, \tag{9}$$

$$\widehat{\mathrm{Var}}[\bar{G}_N] = \int t^2 \mathrm{d}\,\bar{G}_N(t) - \bar{\eta}^2 \; = \frac{\sum \lambda_j}{N} + \frac{\sum\left(\eta_j - \bar{\eta}\right)^2}{N-1}. \tag{10}$$

The finite sample variance of posterior means η_j appears in the second term of equation (10). PMs tend to be under-dispersed because their variance lack the first term, $\sum \lambda_j / N$, from the estimated marginal variance of $\bar{G}_N$. The goal of the CB estimator is to adjust the posterior means to have a variance equal to the estimated marginal variance specified in equation (10). The CB estimate denoted as a_j^{CB} that minimize the posterior expected squared error loss can be defined as follows (Ghosh, 1992):

$$a_j^{CB} = \bar{\eta} + \left(\eta_j - \bar{\eta}\right) \cdot \sqrt{1 + \frac{N^{-1} \sum \lambda_j}{(N-1)^{-1} \sum\left(\eta_j - \bar{\eta}\right)^2}}\;. \tag{11}$$

Since the term in the square root of equation (11) is always positive and larger than 1, the a_j^{CB}'s are more dispersed around $\bar{\eta}$ than are PMs (Shen & Louis, 1998).

The triple-goal estimator aims to obtain a single set of estimates that could satisfy the three inferential goals simultaneously. In essence, however, the triple-goal estimator is designed to minimize the losses for two of the goals: estimating the EDF of τ_j's, G_N, and estimating the rank of τ_j, $\mathbf{R}_j$. The abbreviation GR reflects the two direct inferential targets and is often used to denote the triple-goal estimator in the literature (e.g., Paddock et al., 2006). This chapter also uses GR to denote the estimator.

The GR estimator starts with estimating the posterior mean of the rank of each τ_j, denoted as $\bar{R}_j$, which minimizes the MSELR in equation (6):

$$\bar{R}_j = \mathrm{E}\left[R_j | \hat{\tau}_j\right] = \sum_{k=1}^{N} \Pr(\tau_j \geq \tau_k | \, \hat{\tau}_j)\,. \tag{12}$$

Then the optimized $\bar{R}_j$ are used to obtain integer ranks $\hat{R}_j = \mathrm{rank}(\bar{R}_j)$. $\hat{R}_j$ is the discretized minimum squared error estimate of the rank.

Next, we define a discretized version of $\bar{G}_N$, the EDF estimate that minimizes the ISEL. Shen and Louis (1998) showed that the optimal discrete EDF estimate with at most N mass points is equal to

$$\widehat{U}_r = \bar{G}_N^{-1}\left(\frac{2r-1}{2N}\right), \qquad r = 1, \cdots, N. \tag{13}$$

Then, the GR estimate of τ_j, denoted as a_j^{GR}, can be obtained by estimating the quantile of the distribution of τ_j's evaluated at $\hat{R}_j$ (Lockwood et al., 2018). That is, a_j^{GR} is equal to $\widehat{U}_{\hat{R}_j}$ where $\hat{R}_j = 1, \cdots, N$.

The estimated a_j^{GR}'s are optimal for estimating the EDF of τ_j's and their ranks because it is based on the two discretized minimum squared error estimators, $\widehat{U}_r$ and $\hat{R}_j$. The GR estimator pays no explicit attention to reducing the MSEL of the individual site-specific parameters τ_j. However, Shen and Louis (1998) argued that the GR estimator tends to produce small MSEL for the individual τ_j's because assigning the $\widehat{U}$'s to co-ordinates by a permutation vector **z** to minimize $\sum \left(\widehat{U}_{z_j} - \eta_j\right)^2$ is the same assignments as those aim to minimize $\sum \left(a_j - \tau_j\right)^2$. Still, the focus and strength of the GR estimator lies in the good estimation of the EDF and ranks of τ_j's. Thus, it is an open question whether it performs well in estimating the individual site-specific parameters under various conditions.

1.3.4 Relaxing distributional assumption for the prior G: Dirichlet process mixture

Instead of assuming that the prior distribution G in equation (2) has a known parametric form such as Gaussian, the Dirichlet process (DP) can be used to set a prior on the unknown distribution G, acknowledging uncertainty about its form (Congdon, 2020). The DP prior has two hyperparameters: a base distribution G_0 and a precision parameter α (Antoniak, 1974). A two-stage hierarchical model incorporating the DP prior can be specified as

$$\hat{\tau}_j | \tau_j, \widehat{se}_j^2 \sim N\left(\tau_j, \widehat{se}_j^2\right) \quad j = 1, \ldots, N,$$

$$\tau_j | \tau, \sigma^2 \sim G \equiv \text{DP}(\alpha, G_0) \quad j = 1, \ldots, N. \tag{14}$$

Since this model allows measurement errors ($\widehat{se}_j^2$) on the observed site-level treatment effects at the first-stage, we referred to this model as a Dirichlet process mixture (DPM) model (MacEachern & Muller, 1998; Basu & Chib, 2003). This model can be viewed as *semiparametric* because the first-stage model for $\hat{\tau}_j$ is a parametric model with Gaussian error distribution but the second-stage model for τ_j allows a nonparametric specification with a DP prior.

G_0 provides an initial best guess of the shape of the prior distribution G, which is commonly taken to be a Gaussian distribution in practice. The precision parameter α then controls the degree of shrinkage of G toward G_0. In other words, α determines the extent to which distributions in the sample space partitioned into measurable subsets $G_1, \cdots, G_S$ are divergent from G_0. To understand the role of G_0 and α more intuitively, it is helpful to refer to a form of the induced prior distribution on the site-specific parameter τ_j, so-called the *Polya urn* scheme (West et al., 1994; Dunson et al., 2007):

$$\tau_j | G_0, \alpha, \tau_1, \cdots, \tau_{j-1} \sim \left(\frac{\alpha}{\alpha + j - 1}\right) \cdot G_0 + \left(\frac{1}{\alpha + j - 1}\right) \cdot \sum_{k=1}^{j-1} \delta(\tau_k) , \tag{15}$$

where $\delta(\tau_k)$ denotes a point mass at τ_k. This conditional prior distribution for τ_j is a weighted mixture of the base distribution G_0 and probability masses at the previous site's parameter values, that is, the EDF of $(\tau_1, \cdots, \tau_{j-1})$. In this scheme, the first site's treatment effect τ_1 is drawn from G_0. Then the second site's treatment effect τ_2 is drawn from G_0 with probability of $\alpha/(\alpha + 1)$ or a new empirical distribution $\delta(\tau_1)$ with probability of $1/(\alpha + 1)$. This sampling rule continues, and for the jth site, τ_j is drawn from G_0 with probability proportional to $j - 1$, the number of previous sites which already have realized site-specific parameters, or is sampled from the new empirical distribution of $\sum_{k=1}^{j-1} \delta(\tau_k)$ with probability proportional to α (Gelman et al., 2013).

α can be viewed as a *prior sample size* in some sense (Gelman et al, 2013), as opposed to the sample size of empirical data N. Thus, a huge α value implies an extreme weight on the (prior) base distribution G_0. In that case, the joint distribution of τ_j's tends to be the product of N independent draws from G_0 (Antonelli et al., 2016) and the second-stage model in equation (14) converges to the Rubin (1981) model with a Gaussian prior. On the other hand, a zero α value imposes a null weight on G_0, which leads to the distribution of τ_j being a point mass of $\delta(\tau_1)$. Then, the second-stage model in equation (14) collapses to a model with all sites sharing the common value of τ_1.

Hence, we can infer that α determines the number of distinct values of τ_j, often referred to as the unique number of *clusters* K generated by the DP. The K is not necessarily an exact representation of the number of *mixture components* C (latent subpopulation with substantive meaning) as specified in finite mixture models, but K can be considered as an upper bound of the C (Ishwaran & Zarepour, 2000). The expected number of K is a function of α and N, given by the sum of the weights of G_0 in equation (15) over all N sites:

$$\mathrm{E}(K | G_0, \alpha, N) = \sum_{j=1}^{N} \frac{\alpha}{\alpha + j - 1} . \tag{16}$$

The hyperprior for α plays an essential role in determining the expected number of clusters and therefore in controlling the posterior distribution over clusters. In practice, it is a standard approach to use a $\mathrm{Gamma}(a, b)$ distribution with fixed hyperparameters, the shape parameter a and the rate parameter b, to capture the uncertainty in α (Escobar & West, 1995). A key issue is whether the choice of a and b may have a substantial impact on the posterior distribution of α, and in turn on the clustering behavior. There exist a group of studies arguing that the choice of hyperparameters is less of a concern because the data tend to be quite informative, resulting in a concentrated posterior even with a high variance prior for α (Leslie et al., 2007; Gelman et al., 2013). On the other hand, another group of studies report that estimation or inference can be sensitive to the specific choice of the hyperparameters and in general to the strategies for selecting α (Dorazio et al., 2008; Dorazio, 2009; Paddock et al., 2006; Murugiah & Sweeting, 2012). Our interest is to evaluate the sensitivity under two different options, diffuse and informative DP priors, particularly in the context of recovering the EDF of τ_j's.

The first option is to specify a diffuse Gamma distribution when a priori knowledge on α or K is absent. Antonelli et al. (2016) chose values of a and b such that α is centered between 1 and N with a large variance to assign a priori mass to a wide range of α values. If $N = 50$, for example, we can assign 25 as the mean of the α distribution and 250 as a variance which is ten times the magnitude of the mean. Given these a priori values for the mean and variance of α, we can obtain the corresponding values of $a = 2.5$ and $b = 0.1$ based on the moments of a Gamma distribution: $\mathrm{E}(\alpha|a,b) = a/b$ and $\mathrm{Var}(\alpha|a,b) = a/b^2$.

This study suggests the second option, using χ^2 distribution to construct an informative prior for α. This strategy is based on the probability mass function for the prior distribution of K induced by a $\mathrm{Gamma}(a,b)$ prior for α and the number of sites N (Dorazio, 2009; Antonelli et al., 2016):

$$\Pr(K|N,a,b) = \frac{b^a \cdot S_1(N,K)}{\Gamma(a)} \cdot \int_0^\infty \frac{\alpha^{K+a-1} \cdot \exp(-b\alpha) \cdot \Gamma(\alpha)}{\Gamma(\alpha+N)} d\alpha, \tag{17}$$

where $S_1(N,K)$ is the unsigned Stirling number of the first kind and $K = 1, \cdots, N$. Suppose $\Pr(K)$ that encodes our prior information for the distribution of the expected number of clusters K. We can obtain a solution for a and b by minimizing the discrepancy between the encoded prior $\Pr(K)$ and the α-induced prior $\Pr(K|N,a,b)$, defined by the the following Kullback-Leibler (KL) divergence measure:

$$D_{\mathrm{KL}}(a,b) = \sum_{K=1}^{N} \Pr(K) \cdot \log\left\{\frac{\Pr(K)}{\Pr(K|N,a,b)}\right\}. \tag{18}$$

Dorazio (2009) proposed specifying a and b to be the values for which $\Pr(K|N,a,b)$ most closely matches the discrete uniform distribution to reflect the absence of explicit prior information. This method attempts to mimic a noninformative prior for K.

Our proposal is to take $\Pr(K) \sim \chi^2(\mathrm{df} = u)$ to more intuitively encode our prior knowledge on the expected number of clusters K and its uncertainty. The χ^2 distribution has only one parameter: a positive integer u that specifies the number of degrees of freedom. Our framework is mainly motivated by the feature of the χ^2 distribution that its mean and variance are u and $2u$, respectively. If it is expected that there are approximately five clusters ($K = 5$) and $N = 50$, then one can simply assume that $\Pr(K)$ follows a $\chi^2(5)$ distribution and specify a $\text{Gamma}(a, b)$ that closely matches $\chi^2(5)$ using equations (17) and (18). Panel A of Figure 1.1 shows the result of the numerical analysis based on a grid search algorithm designed to identify the global minimum of the KL divergence measure defined in equation (18). The Gamma distribution with $(a, b) = (1.60, 1.22)$ obtained as the solution that minimizes the KL closely matches the $\chi^2(5)$ distribution as shown in Panel B of Figure 1.1. This strategy is useful for constructing an informative prior for K, particularly when one wants to impose near- zero probabilities beyond a certain threshold ($K = 25$ in the example shown in Figure 1.1) and to be clear about the prior mean and variance of K.

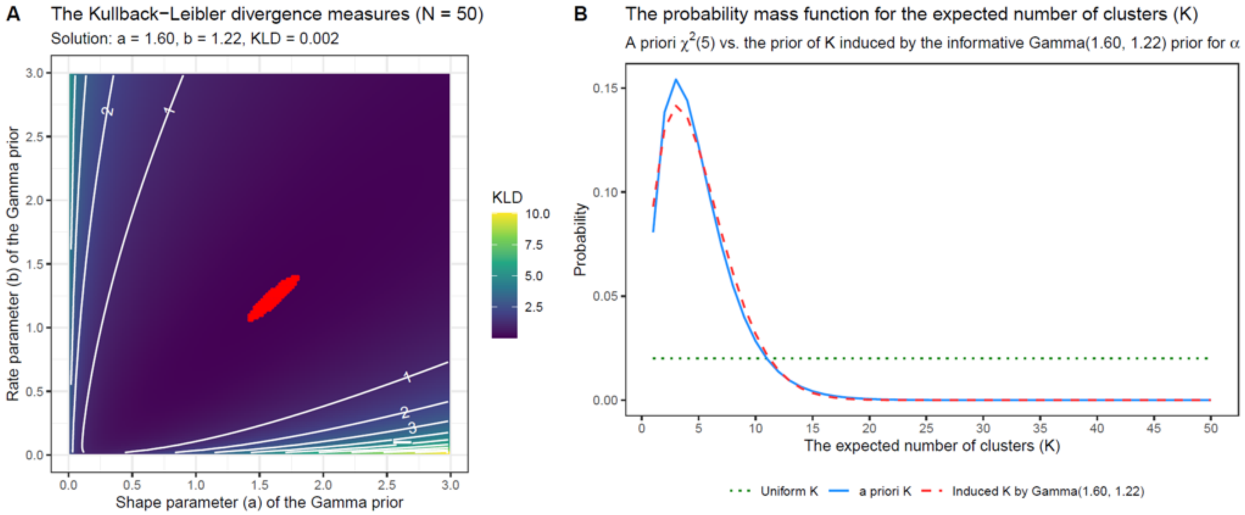

Figure 1.1: The derivation of the informative prior for the precision parameter α by approximating a $\text{Gamma}(a, b)$ to a $\chi^2(5)$

1.4 A simulation study

1.4.1 Design of the simulation

We conduct a comprehensive Monte Carlo (MC) simulation study focusing on relative benefits of the two strategies to improve inferences for site-specific effects: using loss-based posterior summary methods targeted to inferential goals (PM, CB, and GR) and adopting a flexible Dirichlet process for G. We chose to systematically vary four factors for data generation: (a) the number of sites, (b) the reliability of the ML estimates or the shrinkage factor,

(c) the heterogeneity of the $\widehat{se}_j^2$'s, and (d) the true population distribution of G.

The three choices of the number of sites were $N = 25, 50, 200$. Meager (2016) suggests not to use flexible nonparametric specifications in a low-data environment ($N < 50$) because the model risks overfitting the scarce data. Gelman et al. (2013), in contrast, claim that the DP has no tendency for overfitting due to its intrinsic penalty that favors allocation to few clusters that are really needed to fit the data. Our main interest regarding this factor is to evaluate whether increasing N moderates the impact of adopting flexible DP priors on performance measures.

A vector of the first-stage measurement or sampling errors, $\widehat{se}_j^2$'s, is generated by the combination of two factors: the average reliability of the ML estimates $\hat{\tau}_j$ denoted as I and the heterogeneity of the $\widehat{se}_j^2$'s across the N sites denoted as R. Values of I examined in this study are $I = 0.1, 0.5, 0.9$, and those of R are $R = 1, 5, 10$. We denote the resulting vector of simulated $\widehat{se}_j^2$'s as $\widehat{\mathbf{E}}$.

The average reliability I determines how informative the first-stage ML estimates $\hat{\tau}_j$'s are on average. Since we will fix the between-site variance to one ($\sigma^2 = 1$) in all the true data-generating models for G, I is given by $1/(1 + \text{GM}(\widehat{se}_j^2))$ where $\text{GM}(\widehat{se}_j^2)$ represents the geometric mean of $\widehat{se}_j^2$, $\exp\left(\sum_{j=1}^{N} \ln(\widehat{se}_j^2)\right)$. Hence, a large average reliability value indicates less noisy, more informative observed ML estimates $\hat{\tau}_j$'s. If $I = 0.9$, the average within-site sampling variance is about a tenth of the between-site heterogeneity ($\text{GM}(\widehat{se}_j^2) = 0.11, \sigma^2 = 1.00$). We are particularly interested in low informative data environments where $I = 0.1$, which are often encountered in practical applications with small site sizes. If $I = 0.1$, $\text{GM}(\widehat{se}_j^2)$ is nine times as large as σ^2, generating quite noisy $\hat{\tau}_j$'s.

While the I determines the geometric mean of the $\widehat{se}_j^2$'s, R controls the heterogeneity of the $\widehat{se}_j^2$'s across sites. R is defined as the ratio of the largest to smallest $\widehat{se}_j^2$ as in Paddock et al. (2006). The largest and smallest $\widehat{se}_j^2$ can be expressed as a function of I and R: $\widehat{se}_{\max}^2 = R \cdot \left(\frac{1-I}{I}\right)$ and $\widehat{se}_{\min}^2 = \frac{1}{R} \cdot \left(\frac{1-I}{I}\right)$. We can obtain $\widehat{\mathbf{E}}$, a vector of simulated $\widehat{se}_j^2$'s, by taking the exponential of N equally spaced values ranging from $\ln(\widehat{se}_{\min}^2)$ to $\ln(\widehat{se}_{\max}^2)$. The resulting $\widehat{\mathbf{E}}$ reflects both the overall level and heterogeneity of $\widehat{se}_j^2$'s across sites which are encapsulated in the two simulation factors I and R. If $I = 0.1$ and $R = 1$, all site-specific ML estimates will be uniformly noisy. If $I = 0.1$ and $R = 10$, there will be sites that have very noisy ML estimates as well as those with precise estimates, although the overall level of data informativeness is the same as when $I = 0.1$ and $R = 1$.

The true population distribution G is either a Gaussian distribution, a mixture of two Gaussian distributions, or an asymmetric Laplace (AL) distribution. Figure 1.2 shows the shape of each distribution. We consider the Gaussian mixture and AL distributions to examine the non-normal circumstances where true G is multimodal, skewed, or long-tailed. Note that the Gaussian distribution in Panel A of Figure 1.2 has mean 0 and variance 1 ($\tau_j \sim N(0,1)$). For comparability with the Gaussian model, we normalized the Gaussian mixture and AL models to have zero means and unit variances as well. Suppose that the Gaussian mixture data-generating model has two mixture components, $N(\tau_1, \sigma_1^2)$ and $N(\tau_2, \sigma_2^2)$, with a mixture weight for the first component, w. To force this mixture distribution to have mean 0 and variance 1, we define a normalizing factor denoted as C:

$$C = [wu^2 + (1-w) + w(1-w)\delta^2]^{1/2}, \tag{19}$$

where $\delta = \tau_2 - \tau_1$ and $u = \sigma_2^2/\sigma_1^2$. Then, with probability w, τ_j is simulated from the first normalized component $N(-\frac{w\delta}{C}, \frac{1}{C})$, otherwise from the second normalized component $N(\frac{(1-w)\delta}{C}, \frac{u}{C})$ with probability of $1-w$.

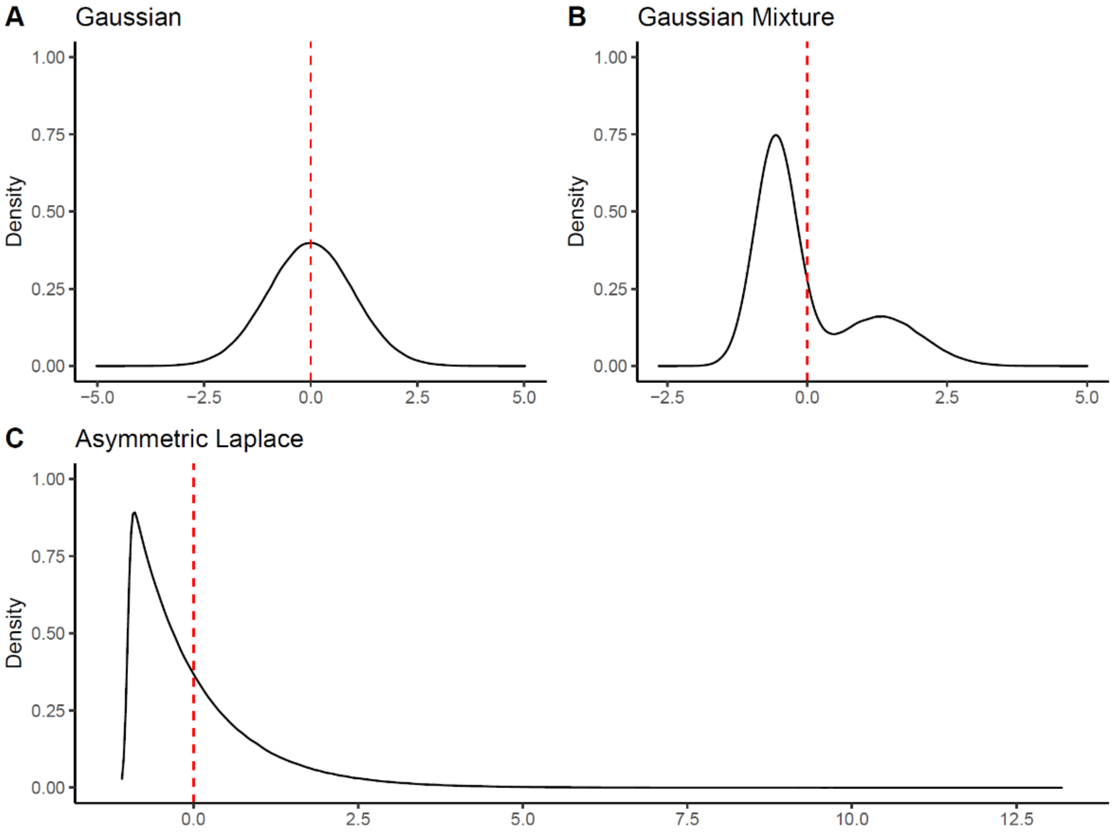

Figure 1.2: The true population distribution G: Gaussian, mixture of two Gaussian, and asymmetric Laplace distributions

To simulate G from the AL distribution with zero mean and unit variance, the location and scale parameters, μ and ψ, are adjusted as follows as functions of the skewness parameter ρ:

$$\tau_j|\mu,\psi,\rho \sim \text{AL}(\mu,\psi,\rho) \quad j = 1,\dots,N,$$

where

$$\mu = -\frac{\psi(1-\rho^2)}{\sqrt{2}\rho}, \quad \psi = \left[\frac{2\rho^2}{1+\rho^4}\right]^{1/2}. \tag{20}$$

The skewness parameter ρ is set to 0.1 so that the resulting G can have a right-skewed distribution with a long tail. This AL data-generating model is useful to evaluate whether the two strategies for the improved τ_j inferences help to recover large but rare site-level effects.

These four factors (N, I, R, and true G) generate $3^4 = 81$ simulation conditions and for each condition 100 datasets are generated. We fit three models to each of the 8,100 datasets: (a) G is standard Gaussian model, (b) G is a Dirichlet process mixture (DPM) model with a diffuse α prior (DP-diffuse), and (c) G is a DPM model with an informative α prior (DP-inform). These three models share the same first stage model specifications: $\hat{\tau}_j|\tau_j, \widehat{se}_j^2 \sim N(\tau_j, \widehat{se}_j^2)$. The models differ based on the second stage specifications for modeling G, $\tau_j|\boldsymbol{\theta} \sim G$ where $\boldsymbol{\theta}$ represents a vector of hyperparameters for G.

The Gaussian model assumes that $G \sim N(\tau,\ \sigma^2)$ where hyperpriors are set to be vague: $\tau \sim N(0,100)$ and $\sigma^2 \sim \text{Unif}(0,100)$. We avoid using the inverse-gamma prior for σ^2 because it does not have any proper limiting posterior distribution particularly for the simulation settings we have, where the number of sites N is small or the site-level variation σ^2 is small by design (Gelman, 2006). The two DPM models assume that $G \sim \text{DP}(\alpha,\ G_0)$ where G_0 is a Gaussian base distribution and α is a precision parameter, and differ in their priors for the precision parameter. Both DPM models, DP-diffuse and DP-inform, share the same base distribution $G_0 \sim N(\tau_j|\tau,\sigma^2)$ with non-informative hyperpriors as in the Gaussian model: $\tau \sim (0,100)$ and $\sigma^{-2} \sim \text{Unif}(0,100)$.

The two DPM models differ depending upon the specification of the $\text{Gamma}(a,b)$ priors for α. In the DP-diffuse models, a and b are chosen so that the mean and variance of α are $\text{E}(\alpha|a,b) = N/2$ and $\text{Var}(\alpha|a,b) = N/5$, respectively. For the three choices of $N = 25, 50, 200$, the pairs of (a,b) for the DP-diffuse model are $(1.25, 0.1)$, $(2.5, 0.1)$, and $(10, 0.1)$. On the other hand, for the DP-inform models, a and b are chosen as the solution that minimizes the KL distance of the distribution of the number of clusters to the χ^2 distribution with $\text{df} = 5$. This means that the DP-inform models assume that the expected number of clusters K is five and that K values larger than about 25 nearly impossible. The obtained solutions were $(1.24, 0.64)$, $(1.60, 1.22)$, and $(1.96, 2.38)$ for the three choices of

$N = 25, 50, 200$. Figure 1.3 shows the distributions of the number of clusters for the two DPM models when $N = 50$. We see that the two DPM models represent contrasting beliefs regarding K. The DP-inform model is more favorable to a more uneven and multimodal G, assuming that about five clusters are needed for 50 sites, while the DP-diffuse model favors smoother G with about 25-30 clusters for 50 sites.

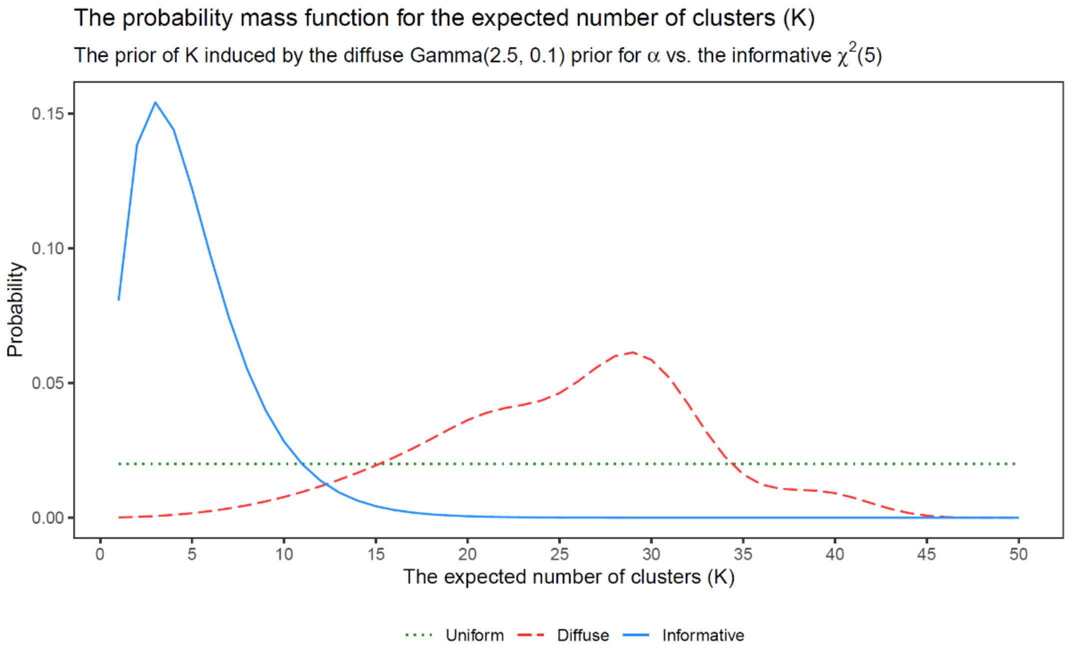

Figure 1.3: The distributions of the expected number of clusters K for the DP-diffuse and DP-inform models ($N = 50$)

For each generated dataset, we use four Markov Chain Monte Carlo (MCMC) chains of length 2,000 with a burn-in of 1,000 for each chain. MCMC simulations generate posterior samples of size 4,000 from the joint posterior distribution of N site-specific effects (we ignore the other parameters). We apply the three posterior summary methods discussed in the previous section to obtain posterior mean (PM), constrained Bayes (CB), and triple-goal (GR) estimates of site-specific effects for each model and dataset.

1.4.2 Performance evaluators

To examine the performance of the strategies for improving inferences for site-specific effects, we mainly assess three performance criteria: (a) the mean squared error loss (MSEL) for the individual site-specific effect parameters τ_j, (b) the mean squared error loss of the ranks (MSELR) based on τ_j, and (c) the integrated squared error loss (ISEL) for the EDF of the τ_j's. These criteria are directly aligned with the inferential goals we illustrated in the previous sections. In addition, we estimate biases in the percentile estimates of G to investigate the performance of the strategies in recovering the unknown density G at specific parts of the distribution. We estimate these biases at the 50th and 90th percentiles. In particular, the discrepancy between the true and estimated 90th percentile estimates reflects the performance of the strategies in recovering large and uncommon impacts at the extreme.

We fit a series of meta-model regressions (Skrondal, 2000) to investigate the relationships between the performance evaluators and the experimental factors. The analysis and reporting of simulation results continues to rely mostly on conventional visual or descriptive analyses (Harwell et al., 2017), although there have been recommendations of using so-called meta-models, model-based inferential analyses of simulation results guided by experimental design (Skrondal, 2000; Boomsma, 2013; Paxton et al., 2001). The meta-models can be useful to accurately detect important patterns and precisely estimate their magnitude.

Our meta-model regressions are fitted separately for the simulation results from three true population distribution G's: a Gaussian, a mixture of two Gaussian, and an AL distributions. An analytic sample from each data-generating model consists of 24,300 observations: a factorial combination of (a) three levels of N (25, 50, and 200), (b) three levels of I (0.1, 0.5, an 0.9), (c) three levels of R (1, 5, and 10), (d) three choices of data-analytic models (Gaussian, DP-diffuse, and DP-inform), (e) three choices of posterior summary estimators (PM, CB, and GR), and (f) 100 replications. To consider the statistical dependence induced by the same data-generating processes, we construct 2,700 cluster (dataset) identifiers (based on N, I, R, and 100 replications), the relevant factors used for data-generation. We denote the total number of observations, 24,300 as n.

Our target outcome variables are five performance evaluators, MSEL, MSELR, ISEL, and biases of the 50th and 90th percentiles. For the explanatory variables, we consider the five simulation design factors, N, I, R, data-analytic models, and posterior summary methods. We construct a set of dummy variables for the five factors with reference groups being $N = 25$, $I = 0.1$, $R = 1$, the Gaussian data-analytic model, and the PM posterior summary method. Accordingly, each design factor with three categories generates two dummy variables, which results in total 10 dummy variables included in the model.

Since the target outcomes are all continuous variables, we adopt a standard linear regression model to specify a meta-model as

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}\,. \tag{21}$$

In this model, $\mathbf{Y}$ is the $n \times 1$ outcome vector and $\boldsymbol{\epsilon}$ is the vector of error terms of the same length. $\mathbf{X}$ is the design matrix that includes (a) a vector of 1's for the intercept, (b) 10 columns that consist of the five times two individual dummy variables for main effects, and (c) 40 columns for two-way interactions of the ten dummy variables. We have 40 two-way interaction variables, not ${}_{10}C_2 = 45$, because the five interaction terms between dummy variables within the same simulation factor are excluded. Hence, the total number of columns of the design matrix $\mathbf{X}$, denoted as p, is 51. $\boldsymbol{\beta}$ is the $p \times 1$ coefficient vector which contains the parameters of interest in our meta-model. We will interpret the meaning of each β coefficient in the next section. Cluster-robust standard errors for $\widehat{\boldsymbol{\beta}}$ based on the sandwich estimator are obtained where the clusters are the 2,700 datasets. This is similar to using a repeated-measures ANOVA where data-analytic model and summary method are *within-subject* factors and the other design variables are *between subject*.

1.4.3 Results

Table 1.1 and Table 1.2 summarize results from the meta-models separately fitted for the simulation results from with the Gaussian mixture and AL distributions. The intercept for each model indicates the estimated mean of loss estimates (MSEL, ISEL, and MSELR) when all simulation design factors are set to be their reference groups: that is, when the Gaussian data-analytic model is used combined with the PM posterior summary estimator under the data-generating condition of $N = 25$, $I = 0.1$, and $R = 1$. This reference condition can be viewed as the most challenging low-data environment where a standard Gaussian-PM approach to modeling site-specific effects is applied. Under this low-data environment it is expected that the Gaussian model with PM estimator produces substantial shrinkage, and hence poor performance for ISEL and MSELR. We observe from the intercepts in both Table 1.1 and Table 1.2 that the loss estimates for the individual site-specific parameters τ_j are far larger compare to those for the EDF of τ_j's and the rank estimates based on τ_j's.

The coefficients for the ten individual dummy variables indicate how much the loss estimates are reduced from the reference condition when a single simulation factor is changed. For example, in Table 1.1, increasing the number of sites N (from 25) to 50 or 200 reduces the MSEL for individual τ_j by 0.236 or 0.410 on average, respectively, which is about 17.6% or 30.6% of the MSEL estimate of the reference condition (1.339). The increased heterogeneity of $\widehat{se}_j^2$ is also helpful to reduce the MSEL for τ_j. When R increases, more individual sites have relatively smaller sampling errors, which leads to lower-error estimates for these lower-error cases (Paddock et al., 2006). Although there are also more sites with higher sampling errors as R increases, it seems that the positive impact of having lower-error sites on reducing the losses overwhelms the negative impact of having higher-error cases under the low-data environment (small N and I).

Table 1.1: Meta-model regression results (Data-generating model: a mixture of two Gaussians)

Outcomes:	MSEL for τ_j		ISEL for the EDF of τ_j's		MSELR for rank	
Terms	β estimate	t	β estimate	t	β estimate	t
Intercept	1,339	44.3	269	41.8	111	100.7
N = 50	-263	-9.8	-38	-6.8	4	3.9
N = 200	-410	-16.5	-82	-17.1	6	6.8
I = 0.5	-747	-29.3	-179	-35.6	-54	-53.3
I = 0.9	-1,186	-47.5	-243	-50.2	-92	-105.2
R = 5	-235	-8.3	-16	-2.9	-2	-2.2
R = 10	-401	-15.9	-55	-11.3	-7	-7.3
DP-diffuse	565	19.4	4	0.6	1	0.7
DP-inform	41	1.5	-3	-0.5	0	0
CB	671	22.8	-107	-20.5	0	0
GR	627	22	-110	-20.8	0	0.3
N = 50 × I = 0.5	197	9.3	17	4.6	-3	-4
N = 200 × I = 0.5	280	14.3	37	11.4	-4	-5
N = 50 × I = 0.9	219	10.6	30	8.2	-3	-3.9
N = 200 × I = 0.9	317	16.6	60	18.8	-7	-10.8
N = 50 × R = 5	43	2.2	-4	-1.1	-2	-3.2
N = 200 × R = 5	80	4.5	3	1.2	-2	-3.1
N = 50 × R = 10	139	7.9	5	1.8	-2	-2.1
N = 200 × R = 10	144	8.9	13	5.1	0	-0.2
N = 50 × DP-diffuse	42	2.3	-3	-0.9	0	-0.6
N = 200 × DP-diffuse	138	8.2	-6	-2.3	0	-0.8
N = 50 × DP-inform	-15	-0.9	-3	-1.1	0	-0.2
N = 200 × DP-inform	-35	-2.2	-6	-2.5	0	-0.5
N = 50 × CB	-52	-3.1	3	1	0	0
N = 200 × CB	-51	-3.3	11	3.8	0	0
N = 50 × GR	-37	-2.3	6	1.9	0	0.2
N = 200 × GR	-23	-1.5	17	6	0	0.1
I = 0.5 × R = 5	271	13.4	10	2.9	5	6.9
I = 0.9 × R = 5	292	14.7	16	4.7	6	9.6
I = 0.5 × R = 10	444	23.8	39	12.3	10	13.3
I = 0.9 × R = 10	473	26	47	15.5	12	19.2
I = 0.5 × DP-diffuse	-721	-37.3	-19	-5.8	0	-0.2
I = 0.9 × DP-diffuse	-738	-39	-13	-4	-1	-1.3
I = 0.5 × DP-inform	-63	-3.6	1	0.5	1	0.7
I = 0.9 × DP-inform	-46	-2.6	8	2.8	0	-0.2
I = 0.5 × CB	-606	-34.5	70	20.2	0	0
I = 0.9 × CB	-679	-39.7	90	26.9	0	0
I = 0.5 × GR	-546	-31.9	72	20.9	0	-0.3
I = 0.9 × GR	-631	-37.9	92	27.6	0	-0.5
R = 5 × DP-diffuse	-84	-4.8	-3	-1	0	-0.4
R = 10 × DP-diffuse	-155	-9.6	-4	-1.7	0	0.6
R = 5 × DP-inform	1	0.1	-1	-0.4	0	-0.6
R = 10 × DP-inform	-3	-0.2	-2	-0.6	0	-0.1
R = 5 × CB	-62	-3.9	3	0.8	0	0
R = 10 × CB	-82	-5.6	7	2.7	0	0
R = 5 × GR	-62	-4	2	0.6	0	-0.4
R = 10 × GR	-74	-5.2	6	2.1	-1	-0.9
DP-diffuse × CB	268	17.6	24	8.2	0	0
DP-inform × CB	18	1.3	1	0.5	0	0
DP-diffuse × GR	256	17.2	17	6	0	-0.3
DP-inform × GR	0	0	-2	-0.7	0	-0.2

Note: The total number of observations $n = 24{,}300$; The β coefficient estimates are multiplied by 1,000.

Table 1.2: Meta-model regression results (Data-generating model: asymmetric Laplace distribution)

Outcomes:	MSEL for τ_j		ISEL for the EDF of τ_j's		MSELR for rank	
Terms	β estimate	t	β estimate	t	β estimate	t
Intercept	1,300	48.1	280	37.7	113	102.2
N = 50	-275	-10.8	-85	-14.7	7	6.5
N = 200	-397	-17.7	-119	-22.4	9	9.8
I = 0.5	-766	-32.6	-186	-33.9	-48	-45.7
I = 0.9	-1,188	-52.4	-247	-46.4	-93	-106.2
R = 5	-201	-8.1	-42	-7.2	-5	-4.9
R = 10	-258	-10.1	-58	-10.2	-4	-3.9
DP-diffuse	578	21.7	20	3.3	0	0.1
DP-inform	57	2.3	2	0.3	0	-0.1
CB	696	25.8	-88	-15.4	0	0
GR	653	24.6	-95	-16.7	0	0.2
N = 50 × I = 0.5	252	11.8	47	12.1	-6	-6.5
N = 200 × I = 0.5	341	17.9	63	18.1	-6	-8.2
N = 50 × I = 0.9	280	13.6	60	16.1	-5	-7.2
N = 200 × I = 0.9	361	19.6	88	26.1	-9	-13.6
N = 50 × R = 5	63	3.5	21	6.1	0	-0.6
N = 200 × R = 5	68	4.4	15	5	0	0.3
N = 50 × R = 10	14	0.7	19	5.9	-3	-3.9
N = 200 × R = 10	27	1.6	15	5.1	-3	-5.2
N = 50 × DP-diffuse	36	2	-1	-0.2	0	-0.3
N = 200 × DP-diffuse	124	7.7	-2	-0.5	0	-0.4
N = 50 × DP-inform	-24	-1.4	-2	-0.8	0	-0.2
N = 200 × DP-inform	-50	-3.3	-3	-1.2	0	-0.4
N = 50 × CB	-64	-3.8	5	1.6	0	0
N = 200 × CB	-61	-4.1	13	4.1	0	0
N = 50 × GR	-40	-2.4	10	3	0	0.2
N = 200 × GR	-25	-1.7	21	6.9	0	0.2
I = 0.5 × R = 5	252	13.4	26	7.5	5	6
I = 0.9 × R = 5	260	14.2	30	8.8	8	12.3
I = 0.5 × R = 10	402	20.8	43	12.8	5	7.2
I = 0.9 × R = 10	425	22.7	49	14.8	12	18.6
I = 0.5 × DP-diffuse	-735	-38.5	-36	-10.8	0	0.2
I = 0.9 × DP-diffuse	-735	-39.9	-29	-8.7	0	-0.7
I = 0.5 × DP-inform	-91	-5	4	1.2	0	0.6
I = 0.9 × DP-inform	-41	-2.4	7	2.3	0	-0.4
I = 0.5 × CB	-634	-36	52	14.7	0	0
I = 0.9 × CB	-696	-41	71	20.7	0	0
I = 0.5 × GR	-564	-32.4	57	16.3	0	-0.3
I = 0.9 × GR	-658	-39.3	75	22.2	0	-0.4
R = 5 × DP-diffuse	-102	-6.3	-6	-2.1	0	0.2
R = 10 × DP-diffuse	-164	-10	-9	-3.2	0	0.3
R = 5 × DP-inform	-16	-1.1	-3	-1.1	0	0.1
R = 10 × DP-inform	-20	-1.3	-4	-1.7	0	0.1
R = 5 × CB	-57	-3.9	3	0.8	0	0
R = 10 × CB	-89	-5.9	3	1	0	0
R = 5 × GR	-45	-3.1	4	1.4	0	-0.6
R = 10 × GR	-67	-4.5	4	1.4	-1	-0.9
DP-diffuse × CB	266	17.8	24	8.2	0	0
DP-inform × CB	15	1.1	-1	-0.2	0	0
DP-diffuse × GR	252	16.9	18	6.1	0	-0.2
DP-inform × GR	-12	-0.9	-4	-1.7	0	-0.1

Note: The total number of observations $n = 24{,}300$; The β coefficient estimates are multiplied by 1,000.

The single most influential factor, however, is the reliability of the ML estimates or the shrinkage factor I. In Table 1.1, when the reliability changes (from 0.1) to 0.5 or 0.9, the MSEL for τ_j decreases on average by 0.747 or 1.186 from the MSEL estimated under the reference condition. These are quite substantial changes, which correspond to proportional reductions of the losses of 55.8% or 88.6%, respectively. The impact of I is also sizable for other loss estimates, the ISEL for the EDF of τ_j's and the MSELR for rank estimates. In particular, the I dummy variables and their associated interaction terms show very high impacts on the MSELR while the other predictors have statistically zero or limited effects on reducing the MSELR. This is not surprising considering the relative noisiness of rank estimates compared to other estimates. Goldstein and Spiegelhalter (1996) pointed out that extremely informative data is needed in order for rankings to be useful. Our meta-model results also suggest that the informativeness of the data (I) is almost the only critical factor that determines the quality of rank estimates.

Figure 1.4 shows how the estimated EDF of τ_j's changes with the level of I. The EDF in Figure 1.4 is estimated for the scenario of $N = 50$, $R = 1$, and G given by a Gaussian mixture distribution. The true population distribution G is presented as a bold line. The first column of Figure 1.4 shows the distribution of the ML estimates, $\hat{\tau}_j$'s. As expected, in the scenario of $I = 0.1$, the distribution of the $\hat{\tau}_j$'s is largely overdispersed due to noise or sampling errors. On the other hand, the distribution of the $\hat{\tau}_j$'s closely approximates the true population distribution G in the scenario of $I = 0.9$ because the individual $\hat{\tau}_j$'s are highly informative. We observe from the third row of Figure 1.4 that the choice of data-analytic model or posterior summary method does not make a significant difference when the $\hat{\tau}_j$'s are highly reliable.

The effects of increasing N or R vary largely depending upon the level of $I,$ the informativeness of the observed data $\hat{\tau}_j$. The coefficients of the interaction terms involving I and either N or R in Table 1.1 and 1.2 capture how the effects of N or R are moderated by the level of I. Figure 1.5 provides a visual representation of the interaction effects. Overall, increasing N is an effective strategy only when the reliability of the ML estimates is low ($I = 0.1$), specifically for the MSEL of τ_j estimated by PM and the ISEL of the EDF estimated by GR. In Figure 1.5, the effects of increasing N from 25 to 200 are significantly larger than those of increasing N from 25 to 50, only when $I = 0.1$. In the scenario of highly informative data ($I = 0.9$), increasing N has minimal (mostly statistically insignificant) effects on the loss estimates. Similar patterns are observed for the effects of R. When $I = 0.5$, the effects of increased R tend to be nonsignificant and smaller in magnitude than the effects of N. Since the data-analytic choice for G matters little across scenarios in Figure 1.5, it is not an effective strategy to specify a flexible DP prior to reduce the losses in estimation under the low-data environment with noisy $\hat{\tau}_j$. Instead, strategies such as increasing the number of sites or including more participants within sites to improve the reliability of ML estimates will be helpful to reduce the losses in estimation.

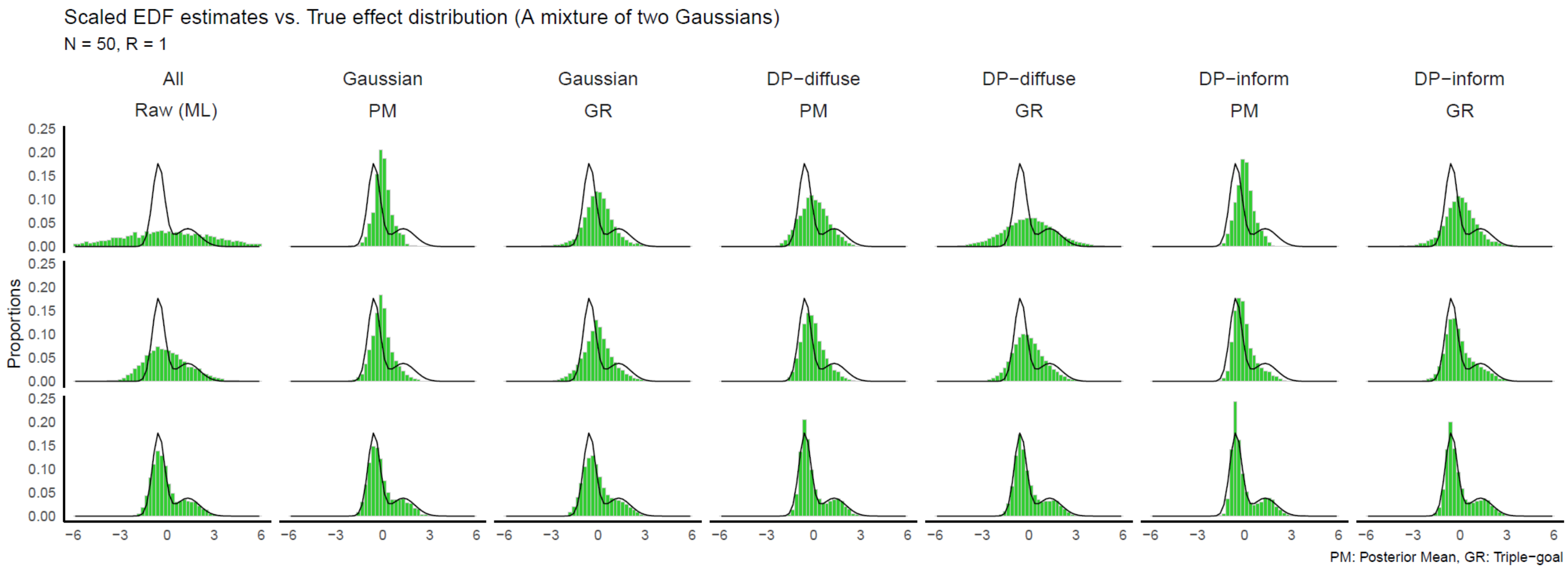

Figure 1.4: Estimated empirical distribution function (EDF) when true population distribution of G is a mixture of two Gaussians.

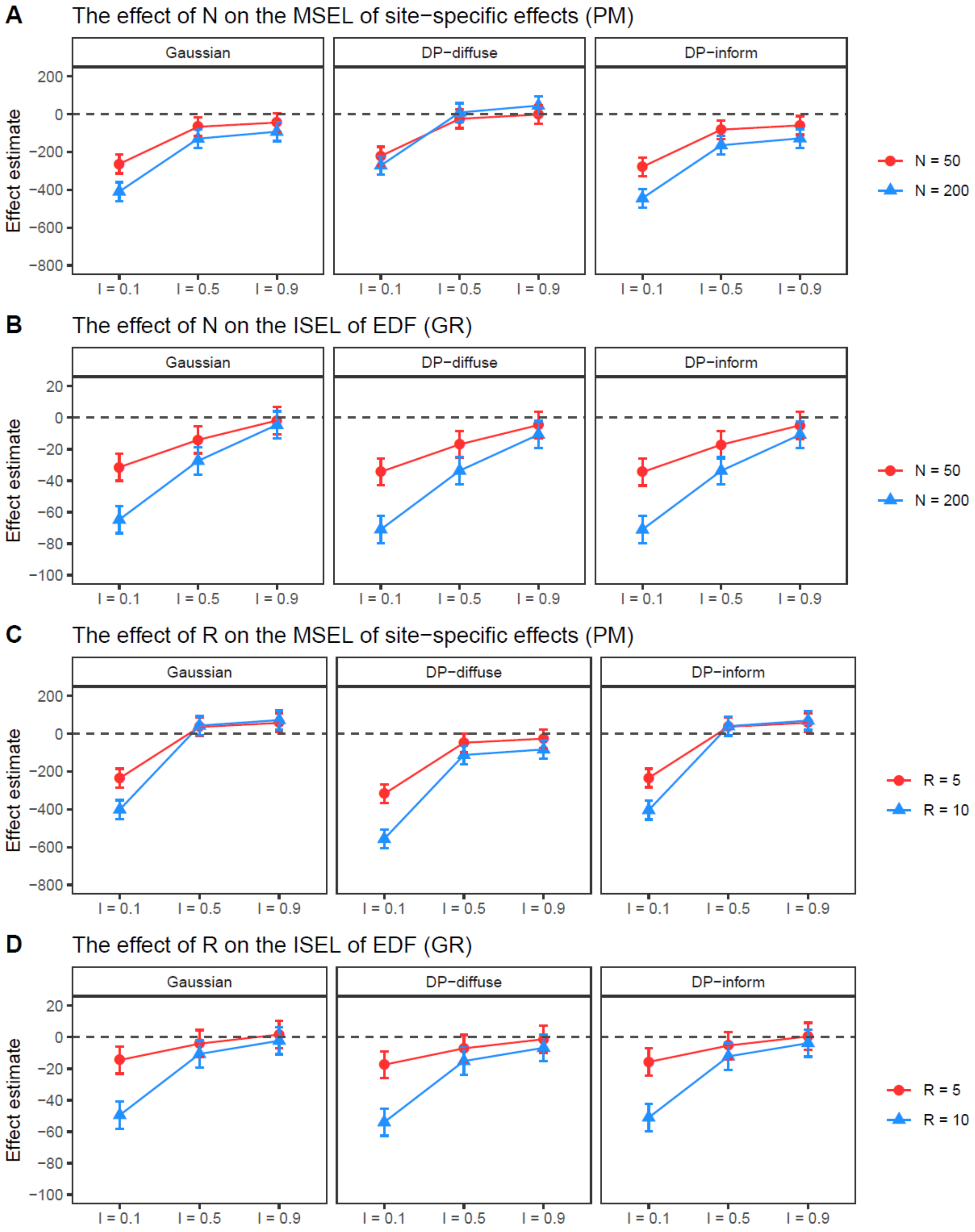

Figure 1.5: The effect of the number of sites (N) and the heterogeneity of $\widehat{se}_j^2$ (R) on the two loss estimates: (a) the MSEL of τ_j estimated using PM and (b) the ISEL of the EDF estimated using GR. The reference group is defined as the scenario where $N = 25,\ R = 1$. True G is the Gaussian mixture model.

We now turn to the impact of the posterior summary method in the low-data environment. In Figure 1.4, the PM estimator tends to yield underdispersed EDF estimates due to the substantial shrinkage to achieve the optimality of the MSEL for individual τ_j. Note in Table 1.1 that using CB or GR estimators rather than PM exacerbates the MSEL for τ_j by 0.671 or 0.627 on average, respectively, increasing the MSEL by about 50% compared with the reference condition (1.339). When CB or GR estimators are used, however, the ISEL for the EDF of τ_j's is reduced on average by 0.107 or 0.110 from the ISEL under the reference condition using PM (0.269). These effects correspond to average proportional reductions of about 40% in the ISEL. Thus, the low-data environment with small I and N requires us to carefully select posterior summary estimators aligned with inferential goals.

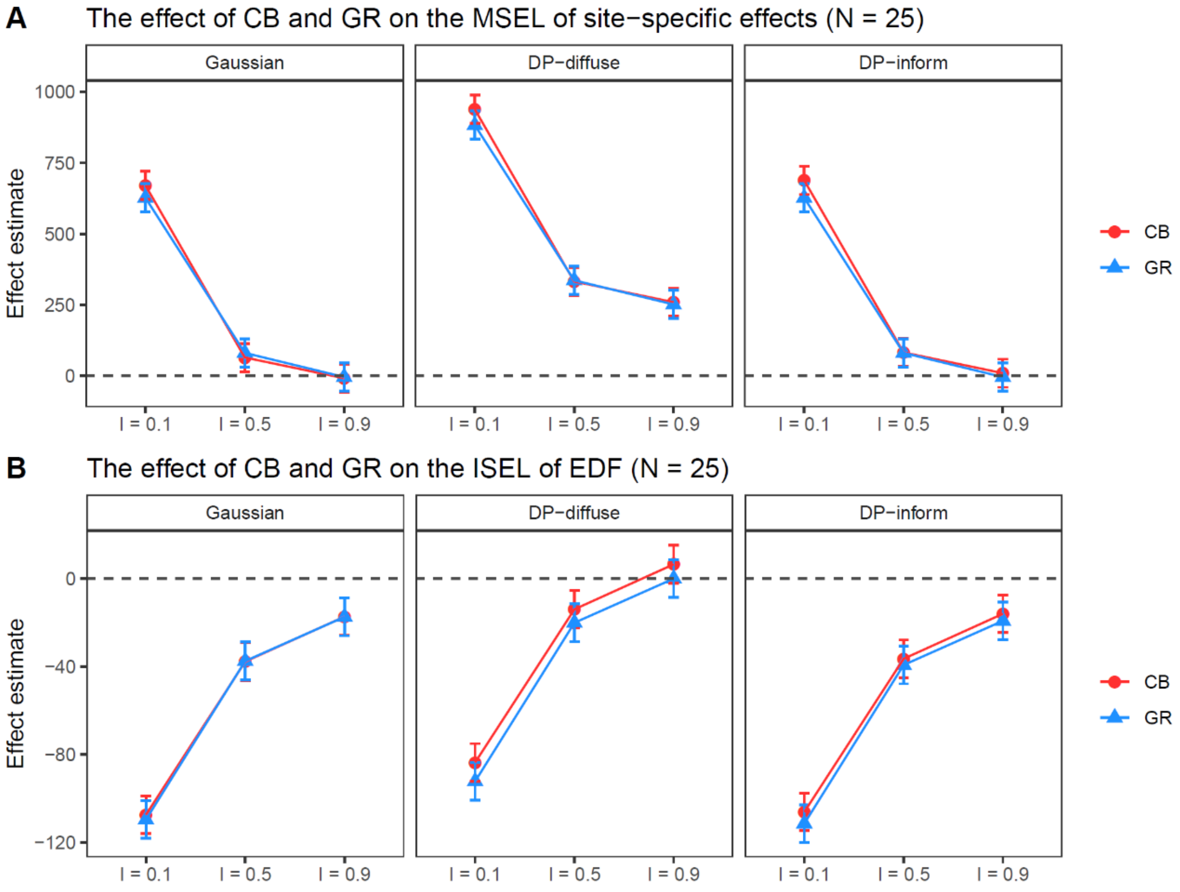

Figure 1.6: The effect of using CB or GR on the two loss estimates: (a) the MSEL of τ_j and (b) the ISEL of the EDF of τ_j's. The reference group is defined as the scenario where $N = 25$, $R = 1$, and PM is used as a posterior summary method. True G is the Gaussian mixture model.

Figure 1.6 shows how the effect of using the CB or GR estimator is moderated by the level of

I and by the data-analytic choice for G in the scenario where $N = 25$, $R = 1$, and true G is a Gaussian mixture. As expected, the performance of CB and GR relative to PM greatly improves with respect to the ISEL of the EDF, while the PM estimates outperform CB and GR in terms of the MSEL of individual τ_j. Yet, the effect of using CB or GR, either positive or negative, decreases in magnitude as the reliability of the ML estimates increases. This indicates that choosing the posterior summary estimator targeting a specific inferential goal is particularly critical in practice when analyzing low-informative data. We also observe that the benefit of using the CB or GR estimator is not significantly different across the data-analytic choice for G, except for the DP-diffuse model that aggravates the performance regarding the MSEL of τ_j. In general, Figure 1.6 suggests that the posterior summary method is a more important factor than the data-analytic choice for G under the low-data environment.

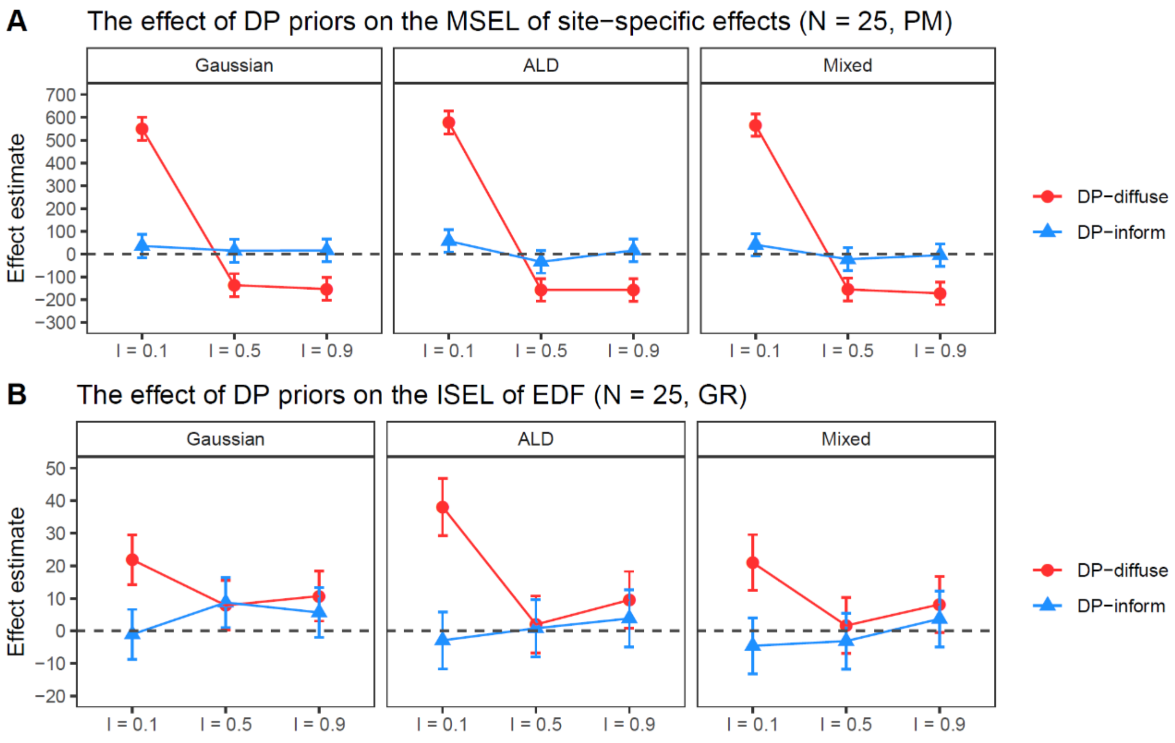

Figure 1.7: The effect of using DP priors on the two loss estimates: (a) the MSEL of τ_j estimated using PM and (b) the ISEL of the EDF estimated using GR. The reference group is defined as the scenario where $N = 25$, $R = 1$, and the Gaussian model is used for G. True G is the Gaussian mixture model.

Figure 1.7 presents how the effect of using DP priors on the loss estimates varies according to the level of I and the true G under the low-data environment. We observe that the two models using different DP priors show quite contrasting loss estimates, which indicates that assumptions on the hyperparameter α can largely affect the posterior distribution when the data are less informative. When $N = 25$, the DP-diffuse and DP-inform models set the prior for α as $\mathrm{Gamma}(1.25, 0.10)$ and $\mathrm{Gamma}(1.24, 0.64)$, respectively. The mean and variance of the first Gamma distribution are 12.5 and 125, while those of the second Gamma distribution are 1.9 and 3.0. Figure 1.4 shows that the EDF estimates of the DP-diffuse model tend to be more dispersed or noisier than those of the DP-inform model mainly because the expected number of clusters K among 25 sites is larger in the DP-diffuse model ($\mathrm{E}(K|G_0, \alpha, N) = 12.1$) than that in the DP-inform model ($\mathrm{E}(K|G_0, \alpha, N) = 5.1$).

The DP-inform model does not make a meaningful difference compared to the Gaussian model under the low-data environment. The DP-inform model tends to produce slightly improved loss estimates for the EDF when the true$G$ is a mixture of two Gaussian distributions because it is more favorable to a multimodal G with a few number of clusters while the DP-diffuse model is more favorable to smoother G with many clusters. Considering the Monte Carlo errors represented as error bars in Figure 1.7, however, this is not a significant benefit of using the DP-inform model relative to the Gaussian models. In contrast, the DP-diffuse model improves the MSLE of individual τ_j across all true G's, only in the scenario where the reliability of the ML estimates is moderate to high ($I = 0.5, 0.9$). But the DP-diffuse model is strongly biased in estimating individual τ_j when $I = 0.1$ and it seems not to be a good choice compared to the Gaussian model in estimating the EDF of τ_j's across all level of I. These patterns are consistent even in the scenario where $N = 50$ or $N = 200$, which suggests that the average benefit of using the flexible DP priors is not considerable relative to the standard Gaussian model in general. Rather, the choice of the posterior summary estimators targeted toward inferential goals has a more decisive impacts on the loss estimates.

Although the DP prior models do not improve the loss estimates on the EDF of τ_j's on average, they may be effective in estimating tail areas of the underlying G, or in general, in estimating percentiles of G. Table 1.3 provides results from the meta-model regression analyses for mean biases in the 50th and 90th percentile estimates of G. We include only the results from the two non-normal true G's, the Gaussian mixture and AL distributions, where the dummy variables and reference conditions are the same as in Tables 1 and 2. Since we observe very similar patterns for the Gaussian mixture and AL distributions, we interpret the estimated coefficient $\hat{\beta}$ for the former here.

Table 1.3: Meta-model regression results for mean biases in the 50th and 90th percentile estimates of G

Outcomes: mean biases	50th Mixture		50th ALD		90th Mixture		90th ALD	
Terms	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t	$\hat{\beta}$	t
Intercept	-19.4	-15.1	-21.6	-16.9	7.6	24.5	6.9	18.9
N = 50	-4.1	-3.5	-0.5	-0.5	2	5.8	2.9	8.4
N = 200	-10.9	-10.6	-10.6	-10.6	3.1	10.6	3.4	11
I = 0.5	2	1.8	4	3.6	-1.1	-3.4	0.4	1.2
I = 0.9	19	18.8	18.3	18.3	-7.5	-23.2	-3.5	-10.7
R = 5	0.5	0.5	2.4	2.1	1.1	3.2	1.2	3.6
R = 10	2.5	2.2	2.2	2	2	6.8	1.7	5.1
DP-diffuse	8.7	7.8	8	7.2	-5.2	-14.3	-5.6	-14.9
DP-inform	-0.4	-0.3	-0.5	-0.5	-0.1	-0.3	-0.7	-2
CB	9.2	8.7	8.7	8.2	-7.7	-21.7	-8.2	-22.2
GR	6.5	6.1	6	5.6	-6.3	-19.7	-7.1	-20.9
N = 50 × I = 0.5	4.1	5	1.4	1.6	-0.8	-2.6	-2.9	-10.1
N = 200 × I = 0.5	10.9	14.8	10.4	14.3	-2.4	-9	-3.5	-13.2
N = 50 × I = 0.9	3.1	4	1	1.3	-0.8	-2.6	-2.8	-10
N = 200 × I = 0.9	7.5	11.1	8	12.1	-1.8	-7	-2.9	-11.4
N = 50 × R = 5	-1.5	-2.1	-2.7	-3.6	-0.4	-1.5	-0.2	-0.8
N = 200 × R = 5	-1.5	-2.3	-1.4	-2.1	-0.7	-2.6	0.1	0.2
N = 50 × R = 10	-3	-4.2	-1.3	-1.8	-1.6	-5.9	-0.1	-0.6
N = 200 × R = 10	-2	-3.1	-0.5	-0.8	-1	-4	-0.8	-3.4
N = 50 × DP-diffuse	1	1.5	1.7	2.5	-0.7	-2.6	-0.2	-0.8
N = 200 × DP-diffuse	3.8	6.1	4.2	6.9	-1.8	-7.3	-1.3	-5.6
N = 50 × DP-inform	1.3	1.8	1.1	1.4	-0.2	-0.9	0.2	0.7
N = 200 × DP-inform	3.3	5.2	2.6	4.1	-0.7	-3.1	0	0.1
N = 50 × CB	1	1.3	0.8	1	0.4	1.5	0.8	3.2
N = 200 × CB	2.2	3.2	2.7	4	0.6	2.3	1	4.1
N = 50 × GR	1.4	1.8	1.2	1.5	-0.1	-0.5	0.4	1.5
N = 200 × GR	3	4.3	3.3	4.9	-0.3	-1.2	0.1	0.5
I = 0.5 × R = 5	2.9	3.8	-0.2	-0.2	-1.7	-6.2	-2.2	-8.4
I = 0.9 × R = 5	-0.1	-0.1	-1.9	-2.8	-1.2	-4.5	-2.2	-8.4
I = 0.5 × R = 10	0	0	-2.4	-3.3	-2.8	-11	-2.4	-9.5
I = 0.9 × R = 10	-3.2	-4.6	-3.1	-4.6	-1.9	-7.4	-2.5	-9.9
I = 0.5 × DP-diffuse	-5	-6.9	-4.9	-6.7	5.2	19.5	5.5	21.1
I = 0.9 × DP-diffuse	-3.1	-4.5	-4.3	-6.3	5.4	20.5	6.1	23.9
I = 0.5 × DP-inform	6.2	8.3	4.2	5.6	-1.3	-5.5	-0.5	-2
I = 0.9 × DP-inform	6.1	8.9	4.1	5.9	-0.9	-3.7	0	0.2
I = 0.5 × CB	-3	-3.8	-2.8	-3.5	3.2	12.5	4.7	19.1
I = 0.9 × CB	-8.1	-10.9	-8.4	-11.5	6.3	24.6	6.9	28.7
I = 0.5 × GR	-3	-3.7	-2.4	-3	3.2	13	4.3	18.5
I = 0.9 × GR	-10.2	-13.7	-9.4	-12.8	6.7	27.6	7	30.3
R = 5 × DP-diffuse	0	0	0.1	0.2	0.8	3	1	4.1
R = 10 × DP-diffuse	-0.5	-0.8	-0.5	-0.8	1.4	5.9	1.5	6.5
R = 5 × DP-inform	0.2	0.4	0.6	0.9	0.1	0.4	0.2	0.8
R = 10 × DP-inform	-0.1	-0.2	-0.2	-0.3	0.1	0.4	0.2	0.9
R = 5 × CB	-0.8	-1.2	-0.2	-0.3	0.7	2.7	0.6	2.4
R = 10 × CB	-0.6	-0.9	0	0.1	0.5	2.1	1	4.4
R = 5 × GR	0.6	0.8	0.9	1.3	0.2	0.7	0.2	0.7
R = 10 × GR	1.7	2.5	2.2	3.2	-0.2	-0.7	0.3	1.3
DP-diffuse × CB	-2	-2.9	-1.8	-2.7	-2.1	-8.9	-2	-8.7
DP-inform × CB	0.3	0.4	1	1.4	-0.4	-1.9	0	-0.2
DP-diffuse × GR	-1.7	-2.5	-1.2	-1.8	-2.3	-10.1	-2.1	-9.7
DP-inform × GR	1.3	1.9	2	2.8	-0.8	-3.6	-0.2	-1.2

Note: The total number of observations $n = 24{,}300$; $\hat{\beta}$ represents the estimates for β

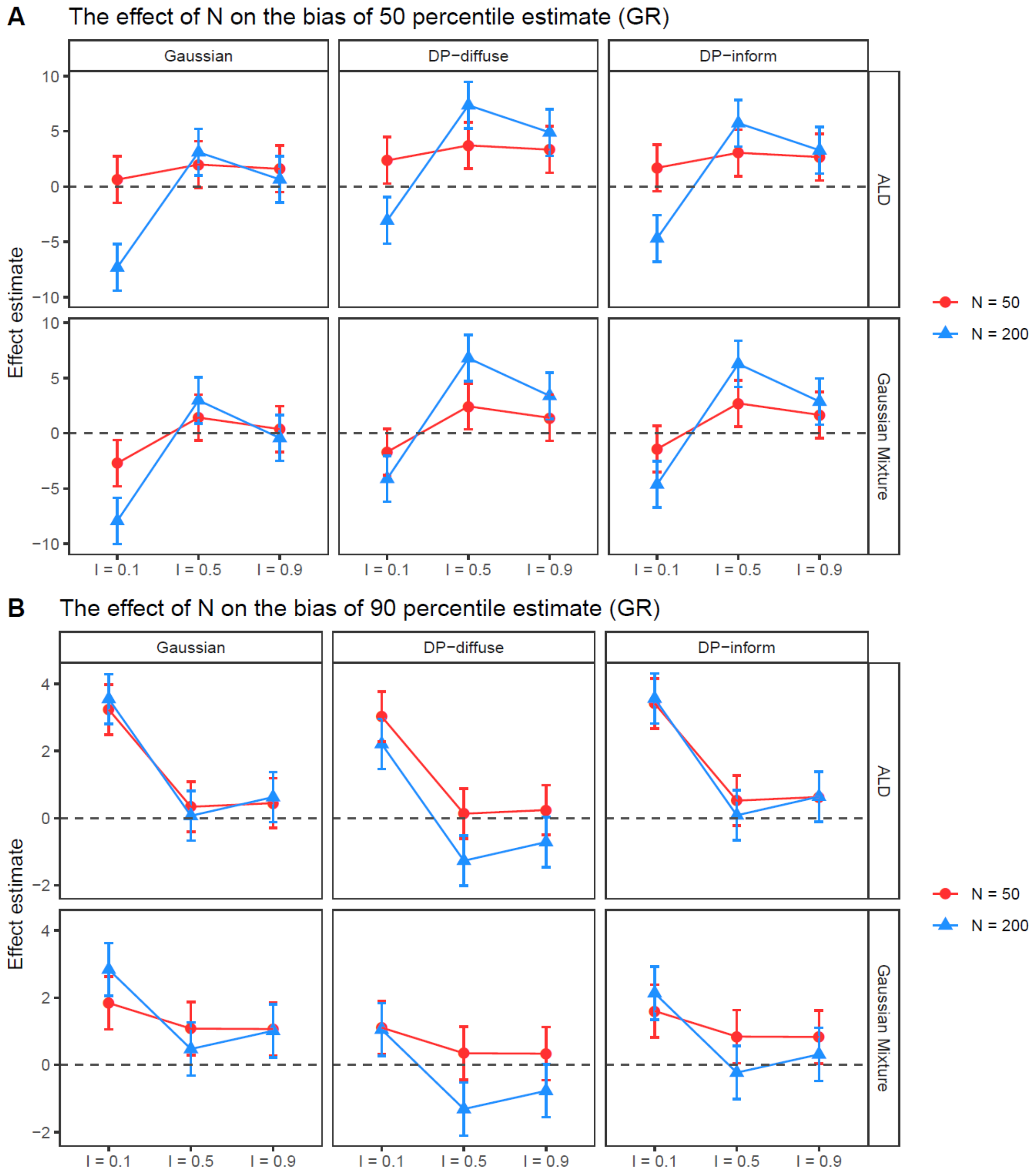

Figure 1.8: The effect of the number of sites (N) on the mean biases of the 50th and 90th percentile estimates of G. The percentiles are estimated by GR. The reference group is defined as the scenario where $N = 25,\ R = 1$.

The mean biases for the 50th and 90th percentiles of G under the reference condition are estimated to be -19.4 and 7.6, respectively. Under the low-data environment, the main source of the biases is shrinkage toward the overall (prior) mean effect. If substantial shrinkage occurs, percentile estimates for the left side of the distribution tend to be underestimated, while those for the right side in the distribution, particularly the right tail or extremity, are overestimated. Hence, positive $\hat{\beta}$'s indicate evidence of bias reduction for the models fitted to the mean biases in the 50th percentile estimates. In the case of the 90th percentile estimates, negative $\hat{\beta}$'s represent effectiveness.

As shown in Table 1.3 and Figure 1.8, increasing the number of sites (N) is not an effective strategy to reduce bias in the percentile estimates when the data is less informative ($I = 0.1$). Instead, when $I = 0.1$, increasing N significantly exacerbates the biases because the distributions of the PM, CB, and GR estimates become more sharply peaked as N increases due to shrinkage. However, when $I = 0.5$ or $I = 0.9$, increasing N improves the percentile estimates based on the DP prior methods. In particular, the DP-diffuse models significantly moderate the effect of increasing N on reducing biases in both the 50th and 90th percentile estimates when the observed ML estimates are (highly) informative.

Figure 1.9 shows that the DP priors are typically effective in reducing biases in percentile estimates of G. When the observed ML estimates are not reliable ($I = 0.1$), the DP-diffuse model is the best choice to reduce biases in percentile estimates. As shown in Figure 1.4, the DP-diffuse model recovers tail areas of G more accurately because they produce smoother and more dispersed PM, CB, and GR estimates than the other models. Panel B of Figure 1.9 suggests that the DP-diffuse model needs to be combined with CB or GR estimators so that it can diminish biases in tail areas of G when analyzing informative data ($I = 0.5$ or $I = 0.9$). On the other hand, the DP-inform model is effective in recovering the true 50th percentile with informative data. This result is related to the DP-inform model's capability to precisely capture the true modes in the distribution (Paddock et al., 2006), which can be seen in the second row of Figure 1.4 where $I = 0.5$.

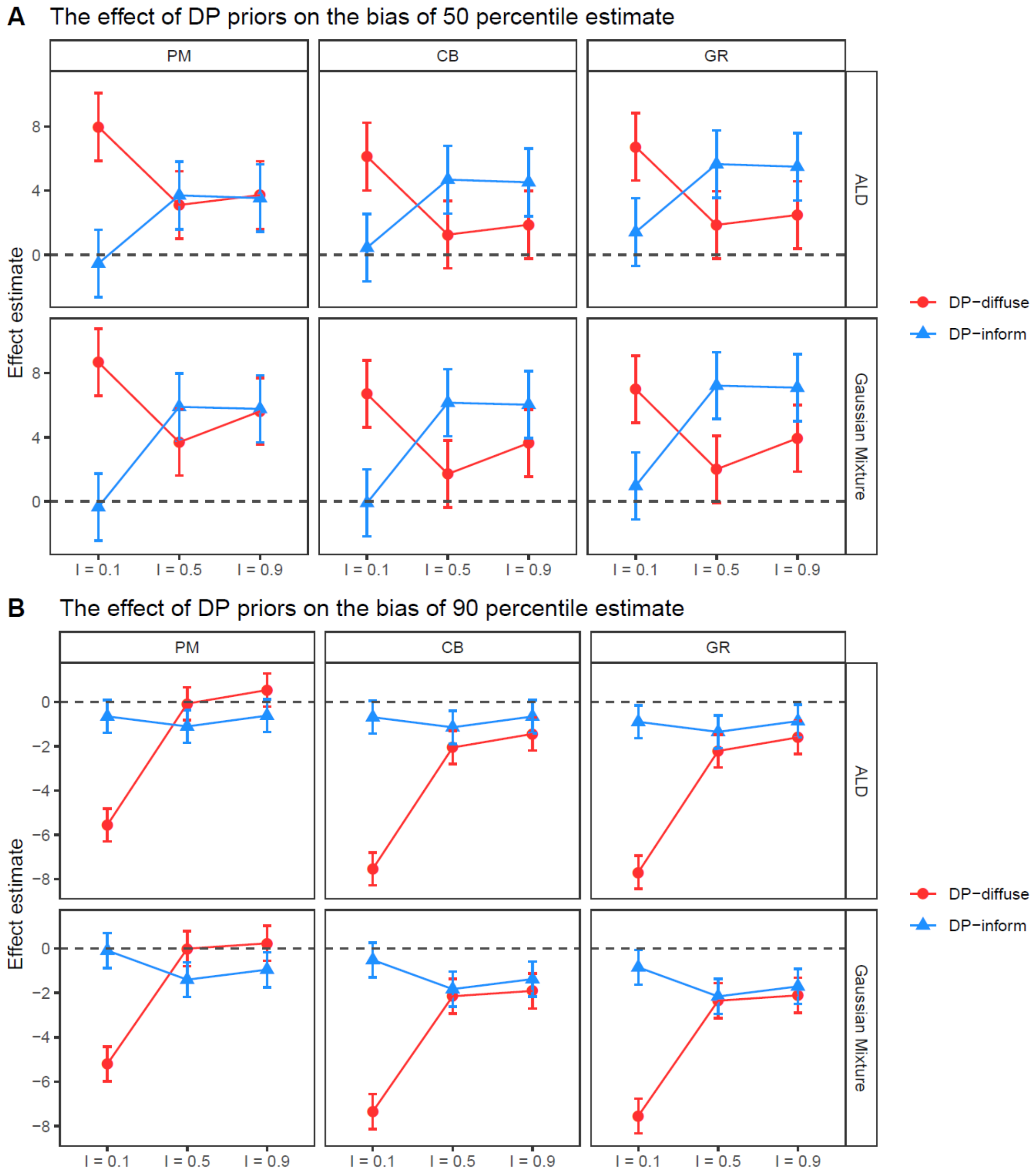

Figure 1.9: The effect of using the DP priors on the mean biases of the 50th and 90th percentile estimates of G. The reference group is defined as the scenario where $N = 25,\ R = 1$.

1.5 Conclusion

Modeling the distribution of site-level treatment effects is an important but formidable task. Among the common inferential goals we considered in this chapter, recovering an unknown prior density G or estimating the EDF of the τ_j's is the most challenging one. We reviewed two strategies to improve inferences regarding τ_j: (a) using posterior summary methods directly targeted toward inferential goals via choice of the appropriate loss functions, and (b) adopting flexible semiparametric specifications for the prior distribution G to protect against model misspecification. We considered the CB and GR for the first strategy, and the DP priors for the second one. The impacts of employing these two strategies on reducing loss estimates were evaluated under varying conditions, particularly focusing on the low-data environment.

The first key finding is that the most influential factor for all inferential goals is the reliability of the ML estimates, that is, the informativeness of the data. This factor was not only the sole predictor that significantly affects all three of loss functions but also was a pivotal *moderator* that determines the effects of other the other simulation factors. Thus, when analyzing data from a multisite trial, the first and foremost task for the analytic decision making is to compute the average reliability of the ML estimates, $I = \sigma^2/(\sigma^2 + \widehat{se}_j^2)$, closely related to the pooling site-specific pooling factor, $\omega_j = \widehat{se}_j^2/(\sigma^2 + \widehat{se}_j^2)$. They provide information on which strategy needs to be employed to reduce the potential losses in the estimation.

When the data are highly informative ($I = 0.9$), the other simulation factors mattered relatively little. This is mainly because the highly reliable ML estimates approximate the true τ_j well, there's little shrinkage in the PM's even with an Gaussian prior distribution, so that the rankings of the τ_j's, as well as their EDF, are recovered well. Accordingly, there is not enough room for the two strategies, CB/GR and DP priors, to improve upon PM with a Gaussian prior. Increasing the number of sites (N) or the heterogeneity of the sampling errors (R) is not helpful either in reducing the losses in the estimation. One can achieve all three inferential goals even with a small number of sites ($N = 25$) if the data are highly informative. This result has an implication for designing a multisite trial – increasing site sizes rather than the number of sites is preferable for the inferential goals based on the site-specific effect parameter τ_j, since between- and within-site variances are out of our control.

When the data are uninformative ($I = 0.1$), on the other hand, study design factors such as N and R as well as the posterior summary estimator significantly affect the performance of estimation. Yet, specifying a flexible DP prior is not in general an effective strategy to reduce the losses in estimation under the low-data environment compared to a Gaussian model. Our findings suggest that the posterior summary estimator matters more and needs to be chosen with great care to align it with the inferential goal. For example, when the inferential goal is to estimate the EDF of the τ_j's with noisy data, the standard Gaussian model with the CB or GR estimators performs better than the flexible DPM models with the PM estimator. In contrast, the standard Gaussian model with the PM shrinkage estimator is more effective than the DPM models with the CB or GR estimators in the estimation of the individual site-specific effect τ_j with uninformative data. Rather than complicating the model specification in a low-data environment ($I = 0.1$), we suggest using a simple parametric model combined with a posterior summary method targeted toward an inferential goal.

Since DPM models are designed to be more adaptive to the data, they require informative data ($I = 0.5, 0.9$) to be effective. Key previous studies recommending the use of DPM models (e.g., Paddock et al., 2006) often consider only moderate to high levels of sampling variation which corresponds to I larger than 0.5. Therefore, our findings are consistent with these previous studies. The DP-diffuse model improves the losses only in the scenario where the data is informative, though less substantially than using appropriate posterior summary methods. However, our findings suggest that the DPM models perform better than the Gaussian model in estimating percentiles of the true underlying G when the data is informative. Combined with the CB or GR estimators, the DP-diffuse models were effective in estimating tail areas of the underlying G because they produce smoother and more dispersed EDF estimates than the other models. On the other hand, the DP-inform model outperforms the Gaussian model when estimating the median because they tend to capture the true modes of the distribution precisely and the median is often close to the mode. Thus, the prior for the precision parameter α needs to be selected with the inferential goal in mind, with different choices preferred for 50th versus 90th percentile estimation.

One of the most important limitations of this study is that site-level covariates are not considered in the data-generating and data-analytic scenarios. Models relying on shrinkage toward the overall mean can be problematic if some specific sites sharing common characteristics vary distinctly in their treatment effects. The multimodality of the site-level effect distribution, for example, might be reflecting such potential moderation of impacts by site-level covariates. In these cases, shrinkage toward a predicted value based on site-level covariates, rather than the overall mean, is more appropriate (Rabe-Hesketh & Skrondal, 2012). Thus, it is another valuable direction of inquiry to investigate how the two strategies to improve inferences for a distribution of τ_j work or fail in scenarios where shrinkage happens misleadingly toward overall mean despite the presence of meaningful site-level moderators.

Chapter 2

Sensitivity of Bayesian Quantile Regression to the Scale Parameter of the Asymmetric Laplace Likelihood

2.1 Introduction

The advantage of quantile regression over conventional linear regression is that it allows a comprehensive analysis of how covariates affect different points of the conditional distributions of the outcome. For instance, in education policy research it can be important to know not just the mean effect of an educational intervention, but also how the effect varies across the achievement distribution. The limitation of estimating only the mean effect has been addressed in the literature by introducing the quantile treatment effect (QTE), the difference between the response's distribution quantiles under the two treatment conditions (Abadie et al., 2002; Bitler et al., 2014; Venturini et al., 2015).

In recent years, the use of Bayesian inference in quantile regression has attracted interest. Bayesian quantile regression methods make use of Markov chain Monte Carlo (MCMC) algorithms to sample the parameter values from their posterior distribution, and thus uncertainty estimates can be calculated easily from posterior samples of MCMC draws. In contrast, frequentist quantile estimators rely on asymptotic methods such as the Wald-type interval estimator (Bassett & Koenker, 1982) or bootstrap. Point estimates can be obtained by minimizing an appropriate loss function through linear programming algorithms without any likelihood function. The usual asymptotic variance-covariance matrix becomes difficult to estimate reliably when the response is censored or when a covariate has missing values (Yang et al., 2015).

Unlike classic quantile regression, Bayesian quantile regression requires a likelihood. There are several proposals including semiparametric likelihoods (Reich et al., 2011; Reich and Smith, 2013), nonparametric likelihoods (Gelfand & Kottas, 2002; Reich et al., 2010), and empirical likelihoods (Otsu, 2008; Yang & He, 2012). By far, however, the most commonly applied likelihood is the asymmetric Laplace (AL) likelihood first employed in Yu and Moyeed (2001). The motivation for using an AL likelihood is that maximizing this likelihood corresponds to minimizing the loss function. Further, Bayesian quantile regression with an AL likelihood is parsimonious, has easily understood interpretations of parameters, and most importantly is computationally convenient for MCMC algorithms.

The fundamental limitation of employing the AL likelihood is that it does not provide a decent approximation to the data generating mechanism. The choice of the AL likelihood is mainly motivated by computational convenience. Therefore, from a modeling point of view, it is a misspecified likelihood. While Yu and Moyeed (2001) argued that the use of AL likelihood is satisfactory in terms of point estimates even if it is not the true underlying distribution and Sriram et al. (2013) established posterior consistency of the point estimators, credible intervals do not generally have correct coverage (Yang et al., 2015; Syring & Martin, 2015). The problem is largely due to the arbitrary scale parameter σ of the AL distribution having a great impact on the variance of the posterior distribution. It is a common practice to fix σ to 1 (e.g. Yu & Moyeed, 2001; Yu & Stander, 2007; Sriram et al., 2013)

Yang et al. (2015) argued that asymptotically correct standard errors can be obtained using a simple adjustment. The proposed adjustment first employs maximum AL likelihood estimation for quantile regression at the median to estimate σ. Then, σ is treated as a known constant, set equal to this estimate, during Bayesian estimation. Finally, an adjustment is applied to the posterior covariance matrix that mimics the *sandwich* estimator, which is borrowed from the frequentist domain.

In spite of having been widely applied in practice, the sensitivity of Bayesian quantile regression to the choice of fixed σ has not been investigated much. Although Yang et al. (2015) provides evidence that interval estimates based on his adjusted standard errors are stable across different values of the fixed scale parameter, their evidence comes from only one empirical data set which is not sufficient for establishing invariance or insensitivity of the posterior inference to the choice of σ. In this article we show that the adjusted uncertainty estimates proposed by Yang et al. (2015) are even more sensitive to the value of σ than the unadjusted posterior standard deviations, which contradicts Yang et al.'s (2015) results. The proposed posterior variance adjustment seems to work only when σ is fixed at the maximum likelihood estimate of the scale parameter at the median. In finite samples, it is shown that point estimates at extreme quantiles can be biased when σ is fixed at very large and unreasonable values.

The outline of the article is as follows. First, we review Bayesian quantile regression methods with an AL likelihood including their computational properties. Then, we compare the non-Bayesian asymptotic variance of estimates with the Bayesian posterior variance with the AL likelihood to discuss the sensitivity of posterior inference to the scale parameter. Finally, sensitivity analyses are performed with simulated datasets and a real data example from education research.

2.2 Bayesian quantile regression with asymmetric Laplace likelihood

2.2.1 Connection between quantile regression and ALD

Let Y be a continuous response variable and and $\mathbf{X}$ a p-dimensional vector of covariates with the first element equal to one. Let $\mathrm{Q}_\tau(Y|\mathbf{X}=\mathbf{x})$ denote the conditional quantile function of Y given $\mathbf{X}=\mathbf{x}$ at a quantile level of $\tau \in (0,1)$. Suppose that the relationship between $\mathrm{Q}_\tau(Y|\mathbf{X}=\mathbf{x})$ and $\mathbf{x}$ can be modeled with the following linear quantile regression model:

$$\mathrm{Q}_\tau(Y|\mathbf{X}=\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}(\tau)$$

where $\boldsymbol{\beta}(\tau)$ is a vector of unknown quantile parameters of interest. Then, consider the working quantile regression model given by

$$y_{\mathrm{i}} = \mathbf{x}_{\boldsymbol{i}}^{\mathrm{T}}\boldsymbol{\beta}(\tau) + \epsilon_i, \qquad i = 1, \dots, n,$$

where ϵ_i is the error term whose distribution is restricted to have the τth quantile equal to zero, that is, $\int_{-\infty}^{0} f_\tau(\epsilon_i)d\epsilon_i = \tau$.

The error density $f_\tau(\epsilon_i)$ often remains unspecified in conventional quantile regression. The unknown parameters $\boldsymbol{\beta}(\tau)$ can be estimated from a random sample of data $D = \{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$ of $(Y, \mathbf{X})$, by minimizing

$$R_n(\boldsymbol{\beta}, D) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\,\boldsymbol{\beta}),$$

where $\rho_\tau(\mu) = \mu\{\tau - I(\mu < 0)\}$ is the quantile loss function and $I(\cdot)$ denotes the indicator function. It is not possible to derive explicit solutions to this minimization problem because the quantile loss function is not differentiable at zero. Thus, linear programming methods are used to obtain quantile regression estimates for $\boldsymbol{\beta}(\tau)$ (Koenker & d'Orey, 1994). In the rest of this section, we denote $\boldsymbol{\beta}(\tau)$ as $\boldsymbol{\beta}$ within equations for simplicity.

A connection between this minimization problem and maximum-likelihood theory is provided by the AL distribution (Yu & Moyeed, 2001; Yu & Zhang, 2005). We say that a random variable y_i is distributed as $\mathrm{AL}(\mu, \sigma, \tau)$ with location parameter μ, scale parameter σ, and skewness parameter τ, if its probability density function is given by

$$f(y_i|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma}\exp\left\{-\frac{\rho_\tau(y_i - \mu)}{\sigma}\right\}.$$

Figure 2.1 shows the AL density for different values of τ. Most of the mass of the ALD lies in the right tail when $\tau = 0.1$ while most of it is situated in the left tail when $\tau = 0.9$. We can see that the skewness of the AL density changes based on the value of τ.

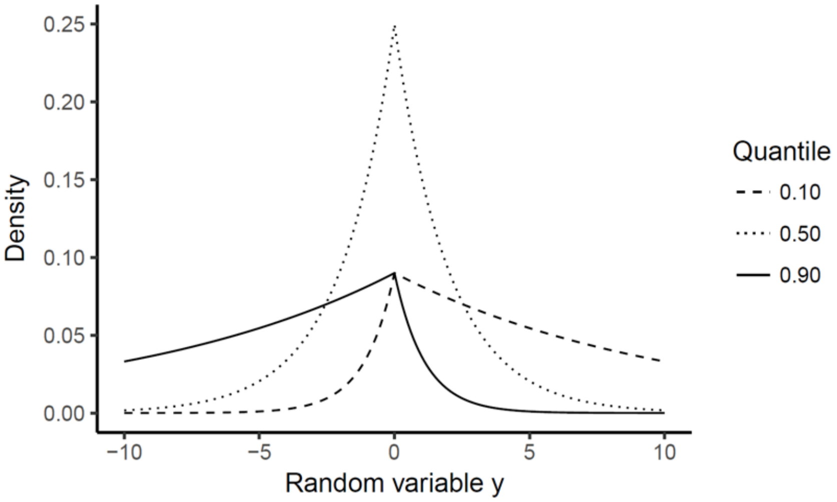

Figure 2.1: Standard ($\mu = 0, \sigma = 1$) asymmetric Laplace densities at $\tau = 0.1, 0.5, 0.9$

By assuming $\mathrm{y_i} \sim AL(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma, \tau)$, the AL likelihood for n independent observations becomes

$$L(\boldsymbol{\beta}, \sigma|D) = \frac{\tau^n(1-\tau)^n}{\sigma^n} \exp\left\{-\frac{\sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\sigma}\right\}.$$

Note that $L(\boldsymbol{\beta}, \sigma|D)$ is proportional to the exponential of minus the scaled loss function $R_n(\boldsymbol{\beta}, D)/\sigma$, with normalizing constant $\tau^n(1-\tau)^n/\sigma^n$. For any value of σ, maximization of the likelihood $L(\boldsymbol{\beta}, \sigma|D)$ with respect to the parameters $\boldsymbol{\beta}$ is equivalent to minimization of the loss function $R_n(\boldsymbol{\beta}, D)$.

The fact that the skewness of the AL density depends on the quantile τ that we happen to be interested in explains why the AL likelihood cannot be the data generating distribution. As shown in Figure 1, the AL density is unimodal and the mode occurs at zero. Since τ determines the skewness of the residual, with low τ implying right-skewed residuals and high τ left-skewed residuals, the AL likelihood can be inflexible in the modeling of the residuals at extreme quantiles, thereby resulting in invalid posterior inference due to model misspecification. It should be noted that the choice of AL likelihood is mainly motivated by computational convenience which will be explained in the next section.

2.2.2 Bayesian computation and properties

Here we outline how the AL likelihood is used in Bayesian quantile regression. With a prior specified as $p_0(\boldsymbol{\beta})$ on $\boldsymbol{\beta}$, the posterior of $\boldsymbol{\beta}$ can be written as

$$p_n(\boldsymbol{\beta}|D) \propto p_0(\boldsymbol{\beta}) \exp\left\{-\frac{\sum_{i=1}^{n}\rho_\tau\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)}{\sigma}\right\}.$$

It has been shown that the use of improper priors for quantile regression parameters $\boldsymbol{\beta}$, including uniform priors, leads to proper posterior distributions of $\boldsymbol{\beta}$ (Yu & Moyeed, 2001; Tsionas, 2003; Choi & Hobert, 2013). Although the computational algorithms may need to adapt to the choice of prior $p_0(\boldsymbol{\beta})$, the asymptotic properties of the posterior are not dependent on the prior choices.

In the Bayesian framework, estimation and inference based on the proposed model can be easily implemented using Markov chain Monte Carlo (MCMC). Although the posterior distribution of $\boldsymbol{\beta}$ is generally intractable, the AL distribution has a nice hierarchical representation which facilitates the implementation of MCMC scheme for Bayesian estimation. The AL distribution can be represented as a scale mixture of normal distribution (Reed & Yu, 2009; Kozumi & Kobayashi, 2011):

Let $y_i \sim AL(\mu, \sigma, \tau)$, $z_i \sim N(0, 1)$, independent of $v_i \sim \exp{(\sigma)}$. Then

$$\mathrm{y_i} \cong \mu + \theta_1 v_i + \theta_2 z_i \sqrt{\sigma v_i},$$

where $\theta_1 = \frac{1-2\tau}{\tau(1-\tau)}$, $\theta_2^2 = \frac{2}{\tau(1-\tau)}$, and $\exp{(\sigma)}$ represents the standard exponential distribution with mean $1/\sigma$, and $\cong$ denotes approximate equality in distribution. Figure 2.2 demonstrates that the scale mixture of normal distribution nicely approximates $y_i \sim AL(\mu = 0, \sigma = 1, \tau = 0.9)$. A Bayesian quantile regression model can therefore be expressed using this hierarchical representation of y_i (Kozumi & Kobayashi, 2011). We simulate draws from the AL distribution by first drawing v_i from an exponential distribution then a normal distribution whose mean is $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \theta_1 v_i$ and whose standard deviation is $\theta_2\sqrt{\sigma v_i}$.

In this posterior sampling scheme, σ can either be fixed or considered as an unknown parameter to be assigned a prior distribution. If σ is assigned an inverse Gamma prior that is independent of $\boldsymbol{\beta}$ as in Kozmi and Kobayashi (2011), the conditional distribution of σ given the other parameters will remain in the inverse Gamma family. Meanwhile, it is a common practice to set σ to a fixed value to make the MCMC algorithm more efficient. For instance, Sriram et al. (2015) fix σ to 1, and Yang et al. (2015) recommend fixing σ at the maximum likelihood estimate of σ at $\tau = 0.5$. The finite sample performance of the posterior, however, is highly sensitive to the choice of σ. We will further discuss this issue in the next section and argue that fixing σ at an arbitrary value for MCMC computation will lead to invalid posterior variance.

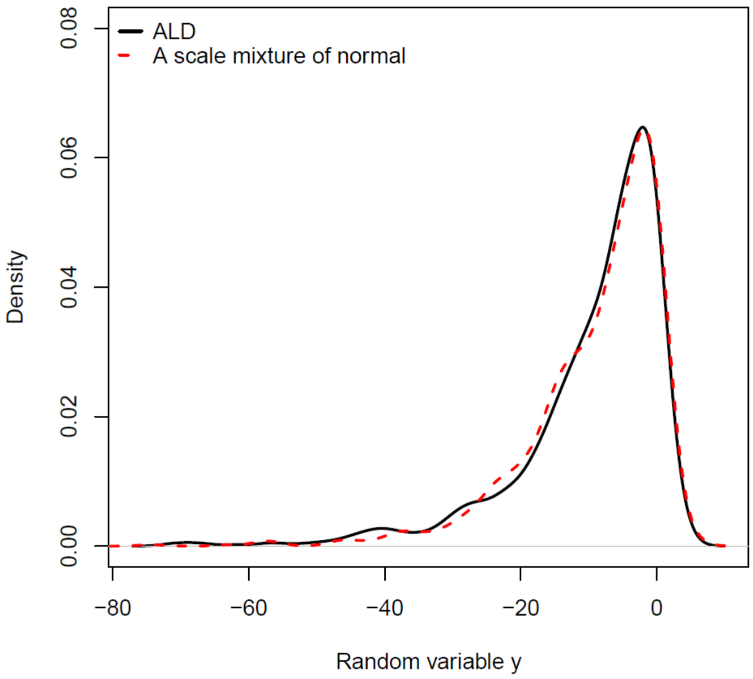

Figure 2.2: Kernel density plot of 1,000 draws from the ALD and a scale mixture of normals with $\mu = 0$, $\sigma = 1$, and $\tau = 0.9$

2.3 Sensitivity of posterior inference to the scale parameter

Generally one of the advantages of a Bayesian approach is that uncertainty of parameter estimates can be estimated directly from the MCMC chains by calculating the posterior standard deviation. Even though earlier literature (Yu & Moyeed, 2001; Sriram et al., 2013) claims that the posterior mean with an AL likelihood is consistent, this does not justify using the uncertainty estimates constructed form the posterior. Recently it has been reported that the stationary distribution for the posterior from Bayesian quantile regression with an AL likelihood does not provide valid posterior inference (Yang et al., 2015; Syring & Martin, 2015). This is mainly because the AL likelihood is misspecified and thus typically yields poorly estimated uncertainty intervals.

We focus on the presence of an extra nuisance parameter σ in the posterior which presents a challenge to valid inference. The role of the scale parameter σ in the AL likelihood is to weight the information in the data relative to information in the prior and to control the spread of the posterior distribution. Consequently, the performance of the posterior is highly sensitive to the choice of this scale parameter. In this section, we explain 1) why the posterior variance-covariance matrix of Bayesian quantile regression with the AL likelihood is invalid and 2) why the matrix is sensitive to the choice of fixed σ. The advantage and limitation of the posterior variance adjustment proposed by Yang et al. (2015) as one way of overcoming the challenges will also be discussed.

2.3.1 The frequentist asymptotic variance of $\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau})$

We start from the established asymptotic validity of the variance of $\widehat{\boldsymbol{\beta}}(\tau)$ in the frequentist domain. Suppose there exist positive definite matrices D_0 and D_1 such that

$$D_0 = \lim_{\mathrm{n}\to\infty} \mathrm{n}^{-1} \sum_{i=1}^{n} \mathbf{x}_\mathrm{i}\mathbf{x}_\mathrm{i}^\mathrm{T}$$

and

$$D_1 = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} f_{y_i}(\,\mathbf{x}_i^\mathrm{T}\boldsymbol{\beta}_\mathbf{0}|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\mathrm{T}.$$

where $f_{y_i}(\cdot\,|\mathbf{x}_i)$ denotes the conditional density of the response y_i given covariate $\mathbf{x}_i$ evaluated at the τth conditional quantile. The analytic methods for obtaining approximate standard errors in the quantile regression model are derived from a general result described in Koenker (2005) giving an asymptotic multivariate normal distribution to the joint distribution of the coefficient estimates $\widehat{\boldsymbol{\beta}}(\tau)$ (Hao & Naiman, 2007):

$$\sqrt{n}\widehat{\boldsymbol{\beta}}(\tau) \overset{d}{\rightarrow} N(\boldsymbol{\beta}_0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}),$$

where the mean of this distribution, $\boldsymbol{\beta}_0$, is the true value of $\boldsymbol{\beta}(\tau)$, and the asymptotic variance, $\tau(1-\tau)D_1^{-1}D_0D_1^{-1}$, takes the form of a Huber-White sandwich estimator (White, 1980).

While the so-called sparsity function D_0 can be estimated easily from the data, estimating the inverse covariance (precision) matrix D_1^{-1} is a challenging task with non-Bayesian methods. Two approaches to the estimation of D_1^{-1} are described in Koenker (2005). One is an extension of sparsity estimation methods suggested by Hendricks and Koenker (1992), and the other is based on kernel density estimation and was proposed by Powell (1986). The key idea of Yang et al.'s (2015) posterior variance adjustment is to use the estimated D_1^{-1} delivered automatically by the MCMC chains in Bayesian quantile regression with the AL likelihood, and to avoid the necessity of estimating the local conditional density $f_{y_i}(\mathbf{x}_i^\mathrm{T}\boldsymbol{\beta}_0|\mathbf{x}_i)$.

2.3.2 The Bayesian posterior variance of $\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ with the AL likelihood

For large n and improper flat prior $p_0(\boldsymbol{\beta}) \propto 1$, we can expect that the posterior density, $p_n(\boldsymbol{\beta}|D)$, is asymptotically a normal density with the following mean and covariance matrix (Yang et al., 2015):

$$p_n(\boldsymbol{\beta}(\tau)|D) \overset{d}{\to} N\left(\boldsymbol{\beta}(\tau), \frac{\sigma D_1^{-1}}{n}\right).$$

The asymptotic posterior variance-covariance matrix from Bayesian quantile regression with the AL likelihood is a function of the precision matrix D_1^{-1}, the fixed scale parameter σ, and sample size n. Since this variance-covariance matrix is not yet adjusted to achieve asymptotic validity, we can denote the variance as $\hat{\Sigma}(\sigma)_{unadj}$.

We can make two claims regarding $\hat{\Sigma}(\sigma)_{unadj}$. First, the asymptotic limit $\sigma D_1^{-1}/n$ is not the right approximation to the sampling variance of $\hat{\boldsymbol{\beta}}(\tau)$ because the asymptotically valid variance-covariance matrix of $\hat{\boldsymbol{\beta}}(\tau)$ was proven to be the sandwhich estimator $\tau(1-\tau)D_1^{-1}D_0D_1^{-1}/n$. This explains why the unadjusted posterior variance is invalid. Second, since the posterior covariance matrix depends on σ, the posterior variance of $\hat{\boldsymbol{\beta}}(\tau)$ is highly likely to be sensitive to the choice of fixed σ. Therefore, setting σ to an arbitrary constant could lead to invalid inferences.

As one way of addressing these challenges, Yang et al. (2015) suggests a simple posterior variance adjustment based on assuming that $\hat{\Sigma}(\sigma)_{unadj}$ is approximately equal to $\sigma D_1^{-1}/n$ and converting that expression to $\tau(1-\tau)D_1^{-1}D_0D_1^{-1}/n$. The adjustment therefore is

$$\hat{\Sigma}(\sigma)_{adj} = \frac{n\tau(1-\tau)}{\sigma^2}\hat{\Sigma}(\sigma)_{unadj}\hat{D}_0\hat{\Sigma}(\sigma)_{unadj},$$

where $\hat{D}_0 = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ provides a consistent estimator of D_0. Yang et al. (2015) argues that the adjusted posterior variance, $\hat{\Sigma}(\sigma)_{adj}$, is asymptotically invariant to the value of σ and yields asymptotically valid posterior inference. They recommend setting σ equal to the maximum likelihood estimate of σ under the AL likelihood at the median, though no proof is provided.

This argument, however, has not been tested much under varying conditions. The posterior density might not approximate the normal density with mean of $\hat{\boldsymbol{\beta}}(\tau)$ and variance-covariance matrix of $\sigma D_1^{-1}/n$, depending upon varying factors such as finite sample size, quantile of interest, value of the scale parameter, and data-generating processes. When the approximation does not work, a specific choice of σ might have some influence on the adjusted covariance matrix. Therefore, it is important to empirically test the invariance or insensitivity of posterior inference to the choice of the fixed σ and other possibly significant factors when using the proposed adjustment method.

2.4 Simulation studies

To conduct a controlled investigation of the sensitivity of Bayesian quantile regression estimates to the scale parameter of the AL likelihood, we conduct a Monte Carlo simulation. We include one covariate x_i which is an independent standard normal variable ($x_i \sim N(0,1)$). The simulation uses 100 replications, and the Bayesian methods use MCMC chains of length 20,000 with a burn-in of 4,000. Data were simulated in R software (R Core Team, 2016), and analyzed using the R package `MHadaptive` (Chivers, 2015) for Bayesian quantile regression and the R package `quantreg` (Koenker, 2016) for conventional quantile regression. We used the Barrodale and Roberts linear programming algorithm for l_1-regression described in Koenker and d'Orey (1994) to estimate the conventional frequentist quantile regression.

2.4.1 Design of the simulation

Two factors were systematically varied for data generation: (a) sample size and (b) data-generating models. For each simulated dataset, three features of estimation were varied: (c) quantile of interest, (d) value of the scale parameter, and (e) uncertainty estimator. We focus on two sample sizes, $n = 200$ and $n = 500$, and three quantiles of interest $\tau = 0.1$, $\tau = 0.5$, and $\tau = 0.9$. For the fixed value of σ, eight values were employed. We first obtain the maximum likelihood estimate of σ under the AL likelihood at the median, $\hat{\sigma}_{\tau=0.5}$, and then obtain the natural log of $\hat{\sigma}_{\tau=0.5}$. By adding eight constant increments $c = \{-4,\ -2,\ -1,\ -0.5,\ 0,\ 0.5,\ 1,\ 2\}$ to $\log \hat{\sigma}_{\tau=0.5}$, we make the magnitude of the fixed scale parameter vary multiplicatively around the $\hat{\sigma}_{\tau=0.5}$, a baseline for comparisons. For example, if we obtained 3.50 for $\hat{\sigma}_{\tau=0.5}$, $\log \hat{\sigma}_{\tau=0.5} + 1$ and $\log \hat{\sigma}_{\tau=0.5} - 1$ would be $\log(3.50 \times e^{1})$ and $\log(3.50 \times e^{-1})$ respectively. Then the set of values for σ will be $\{0.06, 0.47, 1.29, 2.12, 3.50, 5.77, 9.51, 25.86\}$.

Four data-generating models are specified which are similar to models used in Yang et al. (2015)'s simulation studies. For each model, the τth conditional quantile function is $Q_\tau(y_i|x_i) = \alpha(\tau) + \beta(\tau)x_i$, where $\alpha(\tau)$ and $\beta(\tau)$ represents the true quantile coefficients at τ. $\alpha(\tau)$ and $\beta(\tau)$ may or may not vary across quantile levels depending on the data-generating models. When not dependent on the quantile, the fixed parameters are chosen as $(\alpha, \beta, \gamma) = (1.0, 2.0, 0.5)$. γ is a parameter used in Case 3 to model heteroscedastic error.

***Case 1*: True ALD**

$y_i = \alpha + \beta x_i + \epsilon_i(\tau)$, where $\epsilon_i(\tau) \sim AL(\mu = 0, \sigma = 1, \tau)$. The τth conditional quantile function is $Q_\tau(y_i|x_i) = \alpha + \beta x_i$ because $E[\epsilon_i(\tau)|x_i]$ is equal to 0. In this case, the AL likelihood approximates the true data-generating mechanism. Note that different data must be simulated depending on the value of τ which also shows that AL can never be correctly specified if more than one value of τ is considered.

***Case 2*: Location shifted normal**

$\mathrm{y_i} = \alpha + \beta x_i + \epsilon_i$, where $\epsilon_i \sim \mathrm{N}(0,1)$. The τth conditional quantile function is $Q_\tau(y_i|x_i) = [\alpha + \Phi^{-1}(\tau)] + \beta x_i$ where $\Phi^{-1}(\tau)$ is the τth quantile of standard normal distribution. This is a location-shift model in the sense that the intercept is shifted by $\Phi^{-1}(\tau)$, which depends on τ, and a homoscedastic model because the slope $\beta(\tau) = \beta$ is constant across quantile levels.

***Case 3*: Location shifted and scaled normal**

$\mathrm{y_i} = \alpha + \beta x_i + (1 + \gamma x_i)\epsilon_i$, where $\epsilon_i \sim \mathrm{N}(0,1)$. The τth conditional quantile function is $Q_\tau(y_i|x_i) = [\alpha + \Phi^{-1}(\tau)] + [\beta + \gamma\Phi^{-1}(\tau)]x_i$. This model has a location shift of $\Phi^{-1}(\tau)$ and a scale shift of $\gamma\Phi^{-1}(\tau)$. This is a heteroscedastic error model because the x_i has a different effect on different quantiles of the y_i distribution.

***Case 4*: General Location shifted and scaled model**

$\mathrm{y_i} = \alpha(u_i) + \beta(u_i)x_i$, where $u_i \sim \mathrm{Unif}(0,1)$ is independent of x_i, $\alpha(u_i) = \mathrm{sgn}(0.5 - u_i) \cdot \log(1 - 2|0.5 - u_i|)$ and $\beta(u_i) = 2u_i$, where $\mathrm{sgn}(z_i)$ is a sign function used to extract the sign of real number z_i ($\mathrm{sgn}(z_i) = -1$ if $z_i < 0$, $\mathrm{sgn}(z_i) = 0$ if $z_i = 0$, and $\mathrm{sgn}(z_i) = 1$ if $z_i > 0$). The τth conditional quantile function is $Q_\tau(y_i|x_i) = \alpha(\tau) + \beta(\tau)x_i$ for any $\tau \in (0,1)$, where $\alpha(\tau) = \mathrm{sgn}(0.5 - \tau) \cdot \log(1 - 2|0.5 - \tau|)$ and $\beta(\tau) = 2\tau$. This is a location and scale shift model with heteroscedastic error because both intercept and slope depend on the randomly generated quantile level τ.

We consider four different standard error and corresponding interval estimators for $\alpha(\tau)$ and $\beta(\tau)$: for conventional quantile regression, Wald-type standard errors based on asymptotic normality (Bassett & Koenker, 1982) or standard errors based residual bootstrap (De Angelis et al., 1993), and for Bayesian quantile regression with the AL likelihood calculated, both unadjusted and adjusted posterior standard deviations, the latter based on Yang et al.'s (2015) proposal.

2.4.2 Performance evaluators

In evaluating the simulation results, we assess three performance criteria: (a) bias, (b) relative error, and (c) coverage. Suppose that the true parameter is $\beta(\tau)$ and that the ith simulated dataset ($i = 1, \ldots, 100$) yields a point estimate $\hat{\beta}_i$ at the τth quantile with its model-based standard error s_i. We can compute the mean and the variance of *m*=100 estimates of $\hat{\beta}_i$ as $\bar{\beta} = \frac{1}{m}\sum_i \hat{\beta}_i$ and $\mathrm{V}_\beta = \frac{1}{m-1}\sum_i \left(\hat{\beta}_i - \bar{\beta}\right)^2$. We also compute $\overline{s^2} = \frac{1}{m}s_i^2$ and $V_{s^2} = \frac{1}{m-1}\sum_i \left(s_i^2 - \overline{s^2}\right)^2$, because standard theory yields unbiased estimates of the variances, not of the standard deviation (White, 2010). Table 2.1 summarizes how to compute the empirical estimates and their

Monte Carlo (MC) error of the three performance evaluators with these means and variances of $\hat{\beta}_i$ and s_i. The estimated Monte Carlo (MC) error is defined as the standard deviation of each estimated performance evaluator over repeated simulation studies. The MC error estimates represent the uncertainty of the simulation results due to having only a finite number of replicates.

The estimated bias is computed to evaluate the performance of the point estimates $\hat{\alpha}(\tau)$ and $\hat{\beta}(\tau)$. The relative error assesses the performance of the estimated model-based standard error of the $\hat{\alpha}(\tau)$ and $\hat{\beta}(\tau)$. The relative error is based on the ratio of the average model-based standard error ($\sqrt{\overline{s^2}}$) to the standard deviation of the estimates ($\sqrt{\mathrm{V}_{\beta}}$). The latter, often called Monte Carlo standard deviation (MCSD), approximates the true sampling variation. Thus, relative error essentially assesses how well the estimated standard errors represent the true sampling variation. Because we subtract one from the ratio, the relative error will have the value of zero if the mean model-based standard error is equal to the (estimated) true sampling standard deviation, MCSD. Joint performance of the point estimates and their model-based standard errors can be assessed by coverage probabilities. The coverage of a nominal 95% confidence interval is the proportion of 95% confidence intervals that contained the true parameter. Only models that converged properly were included when calculating these three evaluators to cull out improper point estimates and standard errors in the analyses.

Table 2.1: The empirical estimates and their Monte Carlo errors of performance evaluators

Performance evaluator	Estimate	Monte Carlo error
Bias	$\bar{\beta} - \beta$	$\sqrt{\mathrm{V}_{\beta}/m}$
Relative error	$\dfrac{\sqrt{\overline{s^2}}}{\sqrt{\mathrm{V}_{\beta}}} - 1$	$\left(\dfrac{\sqrt{\overline{s^2}}}{\sqrt{\mathrm{V}_{\beta}}} - 1\right)\sqrt{\dfrac{V_{s^2}}{4m\overline{s^2}^2} + \dfrac{1}{2(m-1)}}$
95% coverage probability	$C = \dfrac{1}{n}\sum_i I(\lvert\hat{\beta}_i - \bar{\beta}\rvert < z_{.025}s_i)$	$\sqrt{C(1-C)/m}$

Note: β is the true parameter; $I(\cdot)$ is the indicator function; $z_{.025}$ is the critical value from the normal distribution.

2.4.3 Results

Table 2.2 summarizes the estimates of three performance evaluators and their Monte Carlo errors when the scale parameter is fixed at $\widehat{\sigma}_{\tau=0.5}$ as recommended in Yang et al. (2015). For simplicity of presentation, we first present only Case 1 and 4 with $n = 500$ because the results from Case 2 and 3 have similar patterns as Case 4. Results for Case 2 and 3 can be found in the Appendix.

When the true data generating process follows an AL distribution, we observe that there are no significant biases even for the extreme quantiles ($\tau = 0.1, 0.9$). In contrast, point estimates for the extreme quantiles are systematically biased when the true data generating model is a location shifted and scaled model: $\hat{\alpha}(\tau = 0.1)$ is negatively biased while $\hat{\beta}(\tau = 0.1)$ has positive bias. The intercept and slope parameter estimates at $\tau = 0.9$ demonstrate the opposite directions but similar magnitude of bias. This result confirms previous work (e.g. Cade et al., 2005; Sriram et al., 2013; Hausman et al., 2014) which has shown that the slope parameter estimates of linear quantile regression were biased at extreme quantiles, with lower quantiles biased upwards towards the median-regression coefficients and upper quantiles biased downwards towards the median-regression coefficients. Biases at the extreme quantiles seem not to be due to the misspecification of the likelihood because point estimates from the linear programming method are also biased.

When the data-generating process follows an AL distribution, we see that relative error estimates for the three uncertainty estimators except for the unadjusted Bayesian intervals produce similar patterns. The unadjusted Bayesian uncertainty estimates tend to be greatly inflated at extreme quantiles. The relative error estimates in Case 4 confirm Yang et al.'s (2015) previous observation that the unadjusted Bayesian intervals from the AL likelihood tend to be systematically underestimated which lead to relatively poor coverage. For example, the relative error of the model-based standard error of the point estimate $\hat{\beta}(\tau = 0.5)$ in Case 4 is estimated to be -0.20. This means that the average model-based standard error is estimated to be 20% lower than the MCSD. In contrast, the standard errors based on the proposed variance adjustment is only 4% lower than the MCSD. For $\hat{\beta}(\tau = 0.5)$, the two non-Bayesian methods differ in their performance. While the bootstrapped standard error estimates are similar to the MCSD and adjusted Bayesian standard errors, the Wald-type standard error estimates to be 19% lower than the MCSD. The performance of the two non-Bayesian methods seems to be less stable across quantiles because of the difficulty in approximating the variance–covariance matrices of the quantile estimates.

Table 2.2. Bias, relative error, and coverage probability in case 1 and 4 with $n = 500$ when the scale parameter is fixed at $\widehat{\sigma}_{\tau=0.5}$. MC errors are presented in parentheses.

Data-generating model	Uncertainty estimator	Bias		Relative error		Coverage	
		$\alpha(\tau)$	$\beta(\tau)$	$\alpha(\tau)$	$\beta(\tau)$	$\alpha(\tau)$	$\beta(\tau)$
$\tau = 0.1$							
Case 1: True ALD	$\text{Bayes}_{\text{adj}}$	-.04 (.02)	-.02 (.02)	.19 (.09)	.23 (.09)	.98 (.01)	.99 (.01)
	$\text{Bayes}_{\text{unadj}}$	-.04 (.02)	-.02 (.02)	1.01 (.14)	.99 (.14)	1.00 (.00)	1.00 (.00)
	BR_{boot}	.00 (.01)	-.01 (.02)	.16 (.08)	.10 (.08)	.96 (.02)	.97 (.02)
	BR_{nid}	.00 (.01)	-.01 (.02)	.23 (.09)	.03 (.07)	.97 (.02)	.96 (.02)
Case 4: Location shifted and scaled model	$\text{Bayes}_{\text{adj}}$	-.08 (.01)	.17 (.01)	.15 (.09)	.02 (.08)	.93 (.03)	.62 (.05)
	$\text{Bayes}_{\text{unadj}}$	-.08 (.01)	.17 (.01)	-.18 (.06)	-.13 (.06)	.87 (.03)	.49 (.05)
	BR_{boot}	-.07 (.01)	.16 (.01)	.12 (.08)	.12 (.08)	.94 (.02)	.68 (.05)
	BR_{nid}	-.07 (.01)	.16 (.01)	.14 (.08)	.35 (.10)	.96 (.02)	.82 (.04)
$\tau = 0.5$							
Case 1: True ALD	$\text{Bayes}_{\text{adj}}$	.00 (.01)	-.01 (.01)	.03 (.07)	.13 (.08)	.95 (.02)	.95 (.02)
	$\text{Bayes}_{\text{unadj}}$	.00 (.01)	-.01 (.01)	-.02 (.07)	.06 (.08)	.95 (.02)	.95 (.02)
	BR_{boot}	.00 (.01)	-.01 (.01)	.03 (.07)	.14 (.08)	.97 (.02)	.96 (.02)
	BR_{nid}	.00 (.01)	-.01 (.01)	.08 (.08)	.13 (.08)	.97 (.02)	.96 (.02)
Case 4: Location shifted and scaled model	$\text{Bayes}_{\text{adj}}$	.00 (.01)	.00 (.01)	-.05 (.07)	-.04 (.07)	.88 (.03)	.90 (.03)
	$\text{Bayes}_{\text{unadj}}$	.00 (.01)	.00 (.01)	-.23 (.06)	-.20 (.06)	.85 (.04)	.86 (.03)
	BR_{boot}	.00 (.01)	.00 (.01)	-.05 (.07)	-.05 (.07)	.90 (.03)	.86 (.03)
	BR_{nid}	.00 (.01)	.00 (.01)	-.02 (.07)	-.19 (.06)	.93 (.03)	.86 (.03)
$\tau = 0.9$							
Case 1: True ALD	$\text{Bayes}_{\text{adj}}$	.02 (.02)	.00 (.02)	.17 (.08)	.05 (.08)	.99 (.01)	.94 (.02)
	$\text{Bayes}_{\text{unadj}}$	.02 (.02)	.00 (.02)	.93 (.14)	.71 (.12)	1.00 (.00)	1.00 (.00)
	BR_{boot}	-.02 (.02)	.00 (.02)	.07 (.08)	.00 (.07)	.95 (.02)	.89 (.03)
	BR_{nid}	-.02 (.02)	.00 (.02)	.10 (.08)	-.06 (.07)	.94 (.02)	.91 (.03)
Case 4: Location shifted and scaled model	$\text{Bayes}_{\text{adj}}$	.09 (.01)	-.16 (.01)	.14 (.08)	-.08 (.07)	.89 (.03)	.61 (.05)
	$\text{Bayes}_{\text{unadj}}$	.09 (.01)	-.16 (.01)	-.18 (.06)	-.22 (.06)	.79 (.04)	.57 (.05)
	BR_{boot}	.08 (.01)	-.15 (.01)	.14 (.08)	.00 (.07)	.91 (.03)	.70 (.05)
	BR_{nid}	.08 (.01)	-.15 (.01)	.18 (.09)	.21 (.09)	.93 (.03)	.84 (.04)

Note: $\text{Bayes}_{\text{adj}}$ and $\text{Bayes}_{\text{unadj}}$ represent the Bayesian quantile regression based on an AL likelihood, with and without posterior variance adjustment; BR_{boot} is the frequentist quantile regression using Barrodale and Roberts's linear programming algorithm for l_1-regression with standard error estimates based on residual bootstrapping; BR_{nid} uses the same frequentist model with Wald-type standard errors.

Sensitivity of point estimates to the value of σ.

Figure 2.3 shows estimated biases and their MC errors for the intercept and slope coefficients for each quantile when the AL likelihood is correctly specified. We see that the Bayesian quantile regression estimator yields approximately unbiased slope coefficients, which are not sensitive to the quantile levels and fixed values of the scale parameter. Clearly, the intercept estimates from the Bayesian methods, however, have considerable biases when the fixed scale parameter is large at tail quantiles. As can be seen in Figure 2.4, $\hat{\alpha}(\tau)$ tends to approach the true parameter value when the sample size increases from 200 to 500. Thus, we can expect that biases tend to zero in very large samples, as might be expected given that Sriram et al. (2013) established posterior consistency of Bayesian quantile regression estimator with an AL likelihood. However, we need to note that the finite sample performance of point estimators can be greatly affected by the fixed value of the scale parameter even when the AL likelihood is correctly specified.

Bias estimates when the data generating process follows a location shifted and scaled model are given in Figure 2.5. As we observed in Table 2.2, the intercept estimates are negatively biased and the slope estimates are positively biased at the lower quantile ($\tau = 0.1$) whereas opposite directions of bias are observed at the upper quantile (τ = 0.9). Although biases are present both in conventional and Bayesian quantile regression, the Bayesian estimates are more biased when the scale parameter is fixed at large values. The difference in estimates between the Bayesian and conventional methods is expected to converge to zero as the sample size increases as can be seen in Figure 2.5. We are not sure, however, whether biases themselves at the tail quantiles will be reduced to zero as the sample size increases. In our simulation results, we observe that the intercept parameter tends to be estimated with bias when the quantile regression model is fitted to data generated with location shift (Case 2, 3, 4), and the slope parameter estimates are biased in the scaled data generating models (Case 3, 4). Therefore, researchers working with finite samples need to be aware of the possibility of biased estimates at the extreme quantiles.

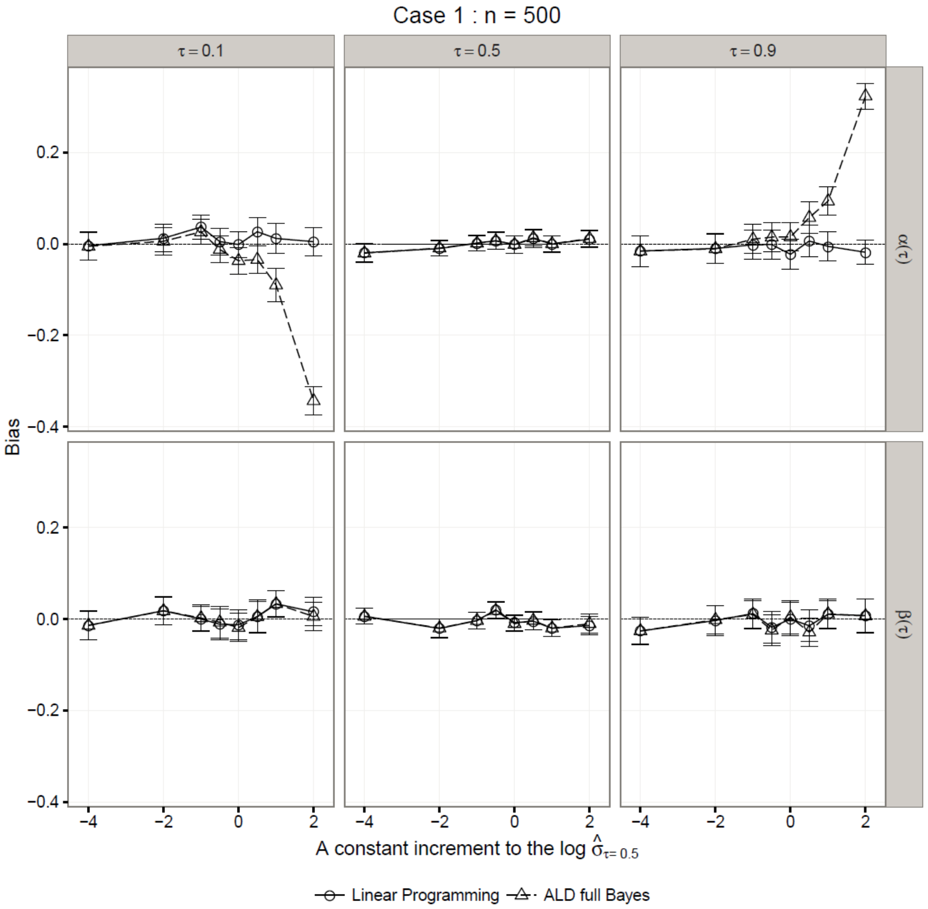

Figure 2.3: Sensitivity of bias to the fixed scale parameter σ in Case 1 with $n = 500$

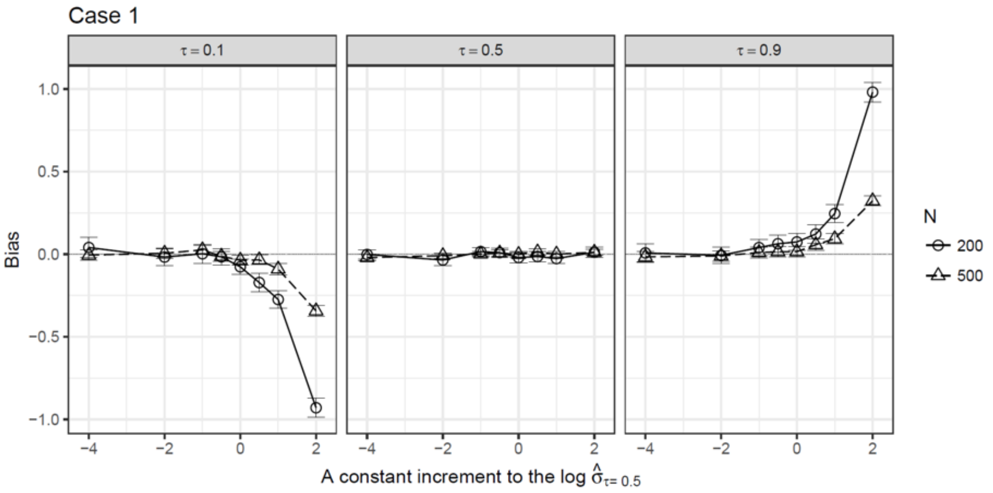

Figure 2.4: Sensitivity of the biases of $\hat{\alpha}(\tau)$ to the fixed scale parameter σ in Case 1 with $n = 200$ and $n = 500$

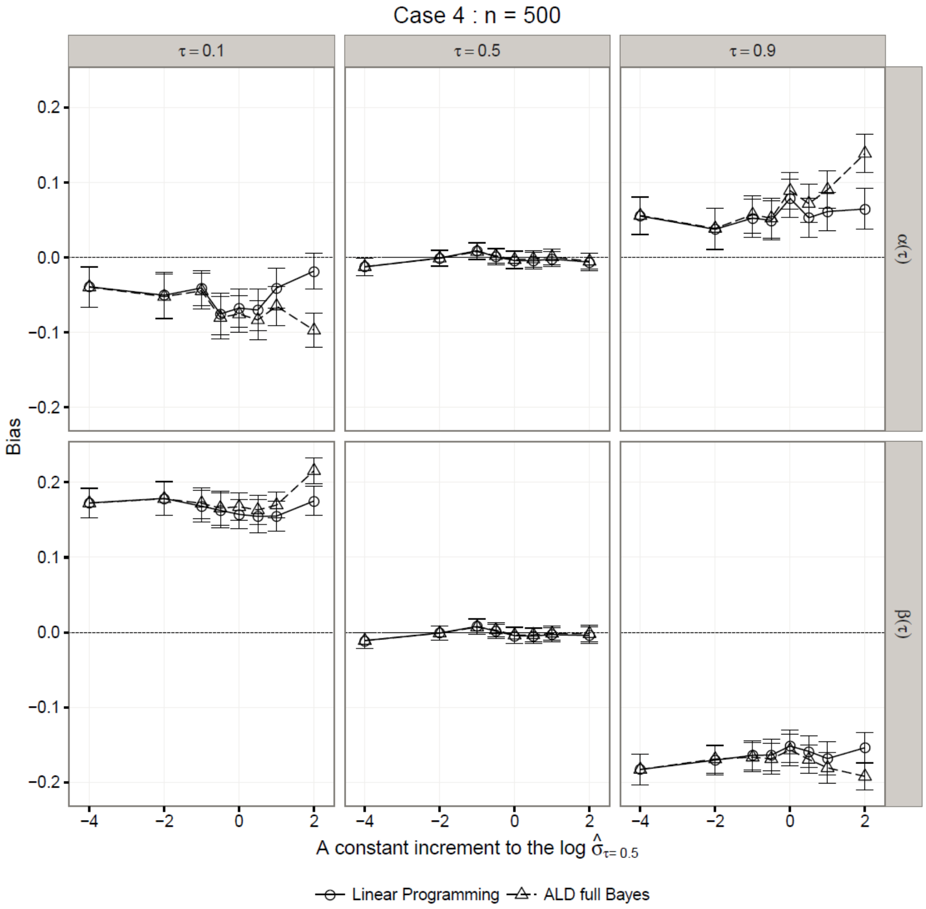

Figure 2.5: Sensitivity of bias to the fixed scale parameter σ in Case 4 with $n = 500$

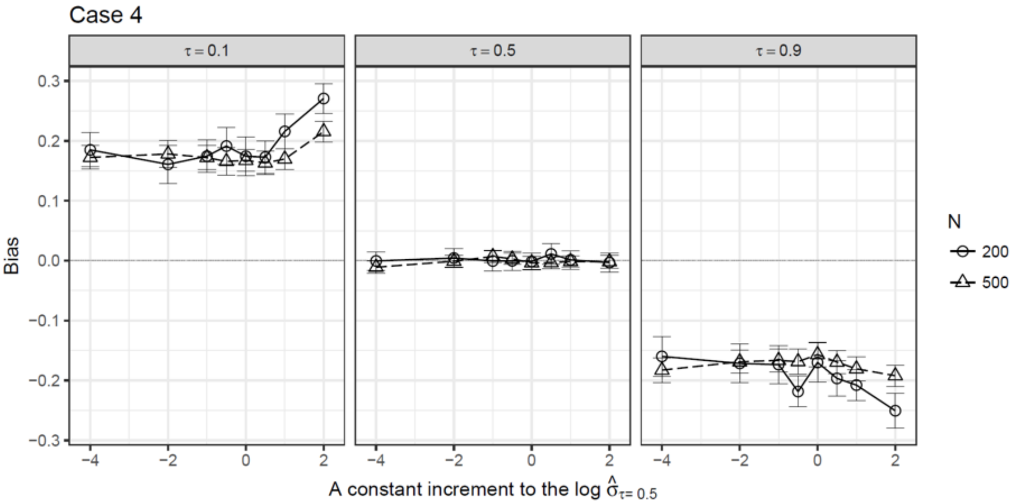

Figure 2.6: Sensitivity of the biases of $\hat{\alpha}(\tau)$ to the fixed scale parameter σ in Case 4 with $n = 200$ and $n = 500$

Sensitivity of model-based standard errors to the value of σ.

Figures 2.7 and 2.8 clearly show that both the unadjusted and adjusted posterior standard deviations depend heavily on the value of σ. As a constant increment is added to the $\log \hat{\sigma}_{\tau=0.5}$, the relative errors of the unadjusted and adjusted posterior standard deviations increase exponentially. It is important to note that the curves for the relative errors of the adjusted and unadjusted standard errors approximately intersect at $\hat{\sigma}_{\tau=0.5}$, (where the x-axis is 0), where the relative error for both methods is close to zero. The only exception is the unadjusted posterior standard deviation estimated at extreme quantiles in Case 1. The unadjusted uncertainty estimates seem to be inflated systematically. We can therefore expect reasonable performance of the Bayesian model-based standard errors if the scale parameter of the AL likelihood is fixed at $\hat{\sigma}_{\tau=0.5}$ as Yang et al. (2015) recommended. However, the standard errors are underestimated if the fixed scale parameter is less than $\hat{\sigma}_{\tau=0.5}$ and overestimated if the scale parameter is fixed at larger values than $\hat{\sigma}_{\tau=0.5}$.

Although Yang et al. (2015) argue that the adjusted posterior variance is asymptotically invariant to the value of σ, the simulation results rather suggest that the adjusted posterior variance is far more sensitive than the unadjusted one in the range around $\hat{\sigma}_{\tau=0.5}$. This pattern persists throughout all the data generating models. Therefore, in order to take advantage of Yang et al.'s (2015) posterior variance adjustment, one should first employ standard frequentist quantile regression at the median of the response on the predictors to estimate the $\hat{\sigma}_{\tau=0.5}$. Setting the scale parameter to an arbitrary fixed constant such as 1 will lead to underestimated uncertainty estimates if the constant is less than $\hat{\sigma}_{\tau=0.5}$.

Sensitivity of the joint performance of point estimates and model-based standard errors to the value of σ.

Since the Bayesian posterior variance increases dramatically as a function of σ, the coverage probability also tends to increase rapidly with the value of σ. When the σ is set to a very large value, overestimated standard errors are therefore expected to lead to coverage close to 1. Thus, we consider only a reasonable range of σ to investigate the joint performance of point estimates and model-based standard errors. Figure 7 plots coverage probabilities just below and above the $\hat{\sigma}_{\tau=0.5}$ for that reason.

Figure 2.9 shows that the Bayesian intervals with adjustment have higher coverage than the corresponding intervals without adjustment in most cases. However, the Wald-type intervals constructed from the conventional quantile regression estimator outperform the two Bayesian counterparts in most cases when the data generating process follows the location shifted and scaled model. The Wald intervals have coverage closer to the nominal 90% level than the Bayesian intervals in most scenarios considered. In contrast, the coverage probabilities of the Bayesian intervals with or without posterior variance adjustment seem less stable because the point estimates and model-based standard errors are sensitive to the choice of σ.

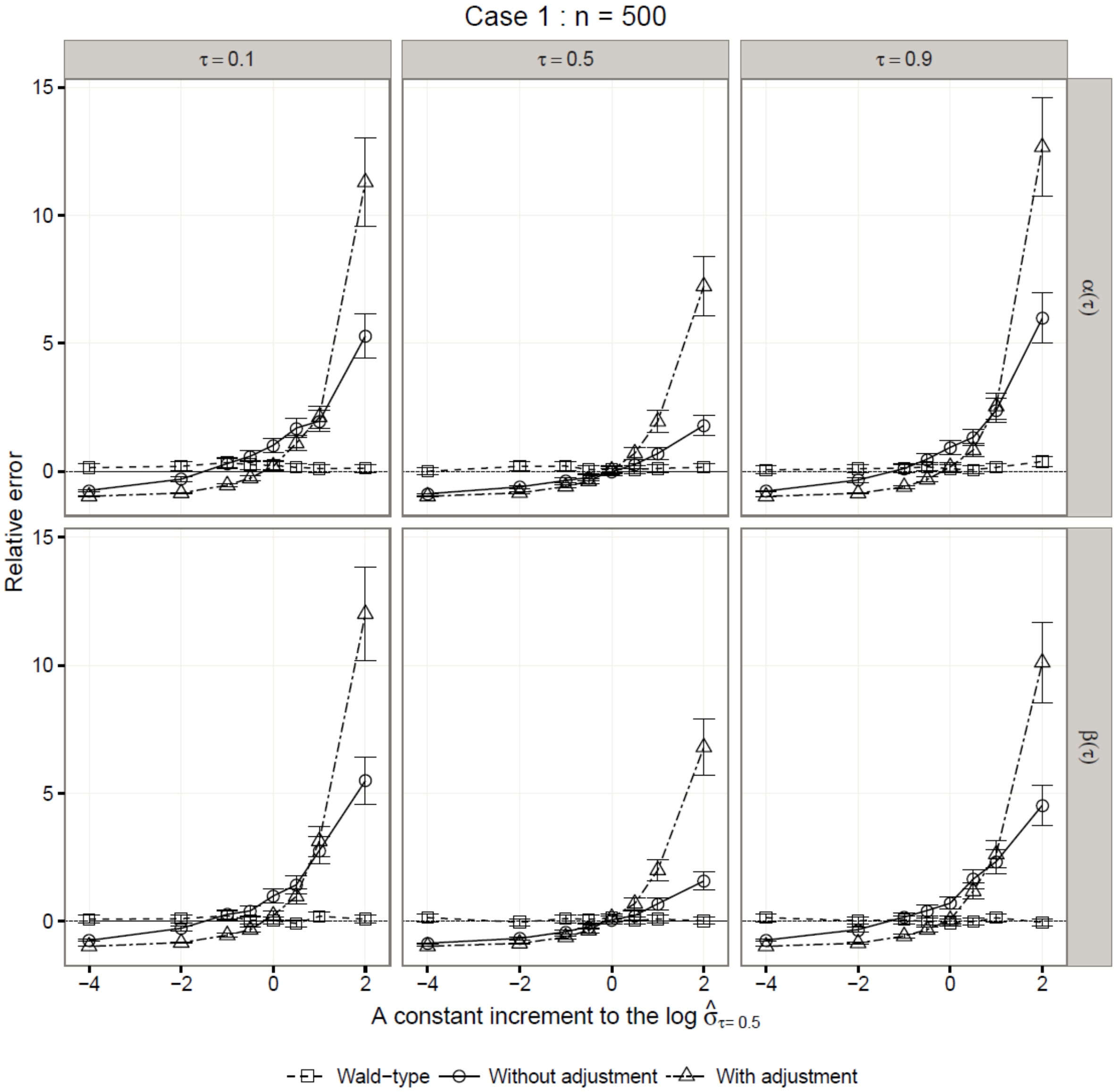

Figure 2.7: Sensitivity of the relative error of the model-based SE to the fixed scale parameter σ in Case 1 with $n = 500$

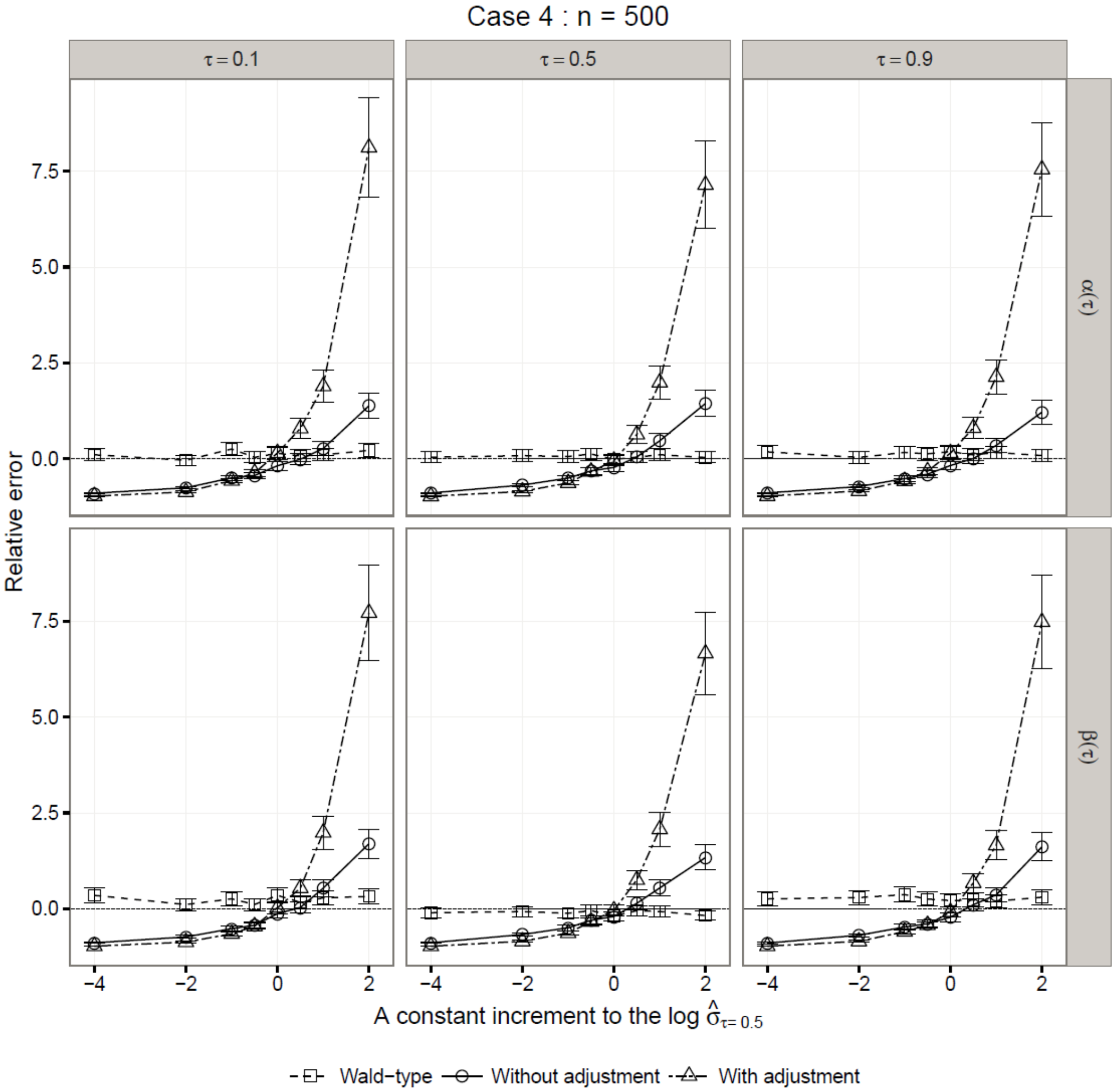

Figure 2.8: Sensitivity of the relative error of the model-based SE to the fixed scale parameter σ in Case 4 with $n = 500$

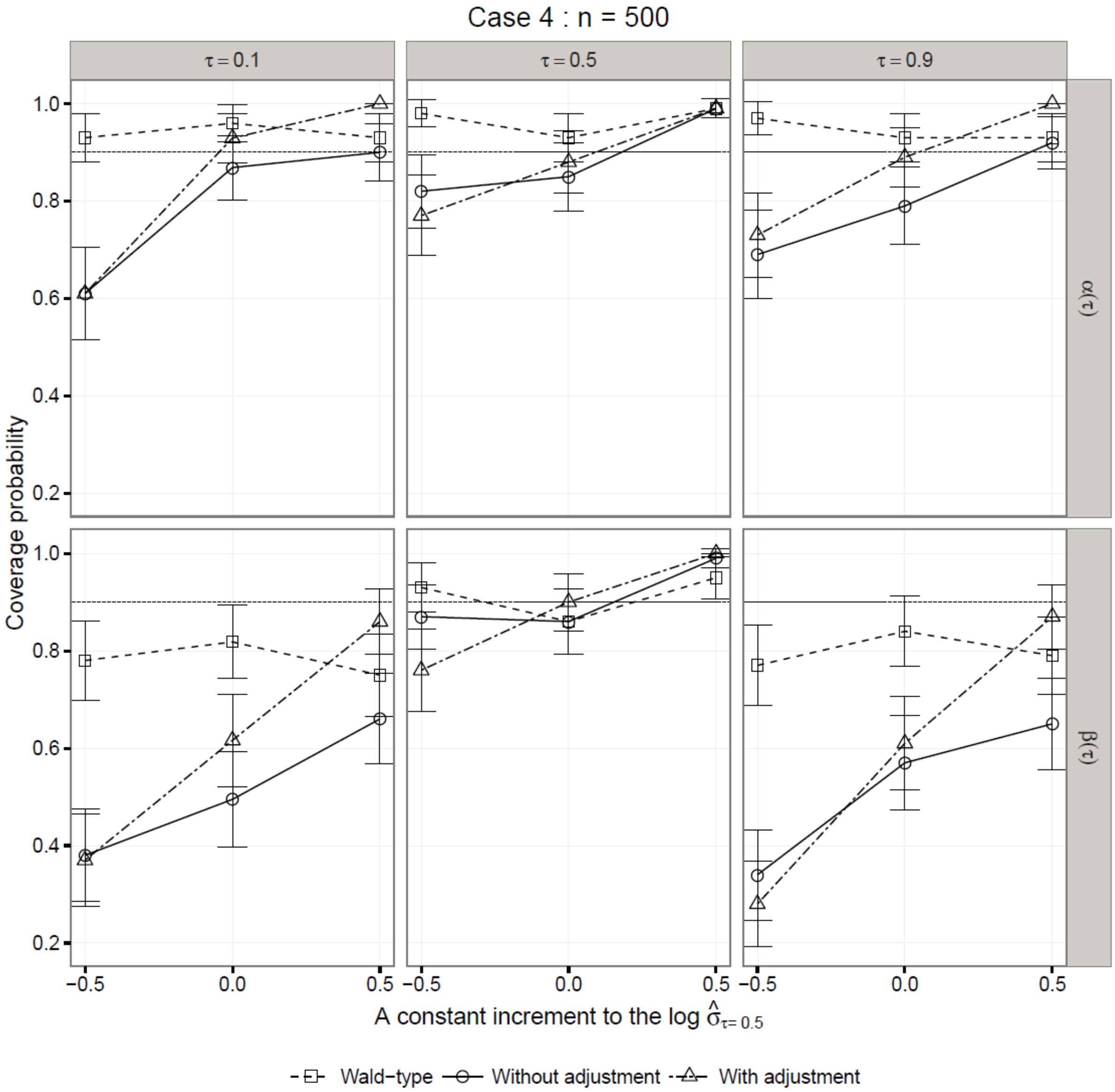

Figure 2.9: Sensitivity of the coverage of 95% confidence/credible intervals to the fixed scale parameter σ in Case 4 with $n = 500$

2.5 Real data analysis

We demonstrate the sensitivity of Bayesian quantile regression to the scale parameter of the AL likelihood using a data from the Programme for the International Assessment of Adult Competencies (PIAAC). PIAAC provides direct measures of cognitive skills such as literacy and numeracy from a nationally representative sample of adults. Taking advantage of these measures, we aim to investigate the relationship between adults' monthly earnings and their literacy skills in South Korea. The outcome variable $lwage_i$ is the natural logarithm of total monthly earnings for salary earner i, expressed in US dollars at purchasing power parity (PPP) rates. The sample mean, minimum and maximum value of the log wage is 8.00, 5.62, and 10.44 respectively. The explanatory variable $literacy_i$ is the z-score of salary earner's literacy skill (with mean of zero and standard deviation one). The estimation sample contains a total of 1,357 adults in South Korea. We consider the following simple linear quantile regression model:

$$\mathrm{Q}_\tau(lwage_i|literacy_i) = \alpha(\tau) + \beta(\tau)literacy_i.$$

This model assumes that wage grows or decreases as a function of the cognitive literacy skills. We focus on quantiles τ = 0.1, 0.5 and 0.9 and estimate $\alpha(\tau)$ and $\beta(\tau)$ using the Bayesian quantile regression estimator based on the AL likelihood with fixed scale parameter σ. We investigate how the point estimates and model-based standard errors of $\alpha(\tau)$ and $\beta(\tau)$ change across varying values of σ. As in the simulation study, we first employ eight levels of the fixed value of the σ by adding eight constant increments $c = \{-4,\ -2,\ -1,\ -0.5,\ 0,\ 0.5,\ 1,\ 2\}$ to the $\log\widehat{\sigma}_{\tau=0.5}$. As $\widehat{\sigma}_{\tau=0.5}$ is estimated to be 0.229 from the data, the set of the varying quantities of the pre-estimated σ will be $\{0.004, 0.030, 0.084, 0.139, 0.229, 0.378, 0.622, 1.692\}$. Then we add larger constant increments such as 4 and 6 to investigate the behavior of point estimates and their uncertainty estimates when the scale parameter is fixed at extremely large and unreasonable values. The scale parameter σ will be set to 12.503 when 4 is added to $\log\widehat{\sigma}_{\tau=0.5}$, and 92.385 when 6 is added to the value.

Figure 2.10 shows 95% credible intervals of the coefficient of literacy as a function of the scale parameter. Similar to the simulation results, this figure suggests that the posterior standard deviation depends heavily upon the fixed value of σ. The two Bayesian intervals are not far from the Wald-type interval when the scale parameter is fixed at $\widehat{\sigma}_{\tau=0.5}$. However, setting the fixed σ to smaller values than $\widehat{\sigma}_{\tau=0.5}$ results in underestimation of the Bayesian uncertainty intervals while setting it to larger values than $\widehat{\sigma}_{\tau=0.5}$ leads to the overestimation of the intervals. It is important to note that the width of the 95% intervals is more sensitive to different values of σ when the posterior variance adjustment is used. In fact, when $\sigma = 1.692$ ($\log\widehat{\sigma}_{\tau=0.5}$+ 2), the posterior variance adjustment yields remarkably larger standard errors, which are about three times of the standard errors estimated without adjustment. This result contradicts Yang et al.'s

(2015) finding that the Bayeisan intervals are stable across different values of σ with the proposed posterior variance adjustment.

Table 2.3 presents the estimates and standard errors and shows a similar pattern as the simulation results. At the lower quantile ($\tau = 0.1$), we see that the $\alpha(\tau)$ is estimated to be smaller with the Bayesian method than with the conventional quantile regression when σ is large, whereas the Bayesian $\beta(\tau)$ estimates are larger than the conventional estimates. Opposite directions of biases are observed at the upper quantile ($\tau = 0.9$). Although we are not sure whether or not the conventional estimates are biased, it seems evident that the Bayesian point estimates deviate more from the conventional estimates as σ increases. This deviation can be regarded as biases introduced by fixing the scale parameter at very large and unreasonable values.

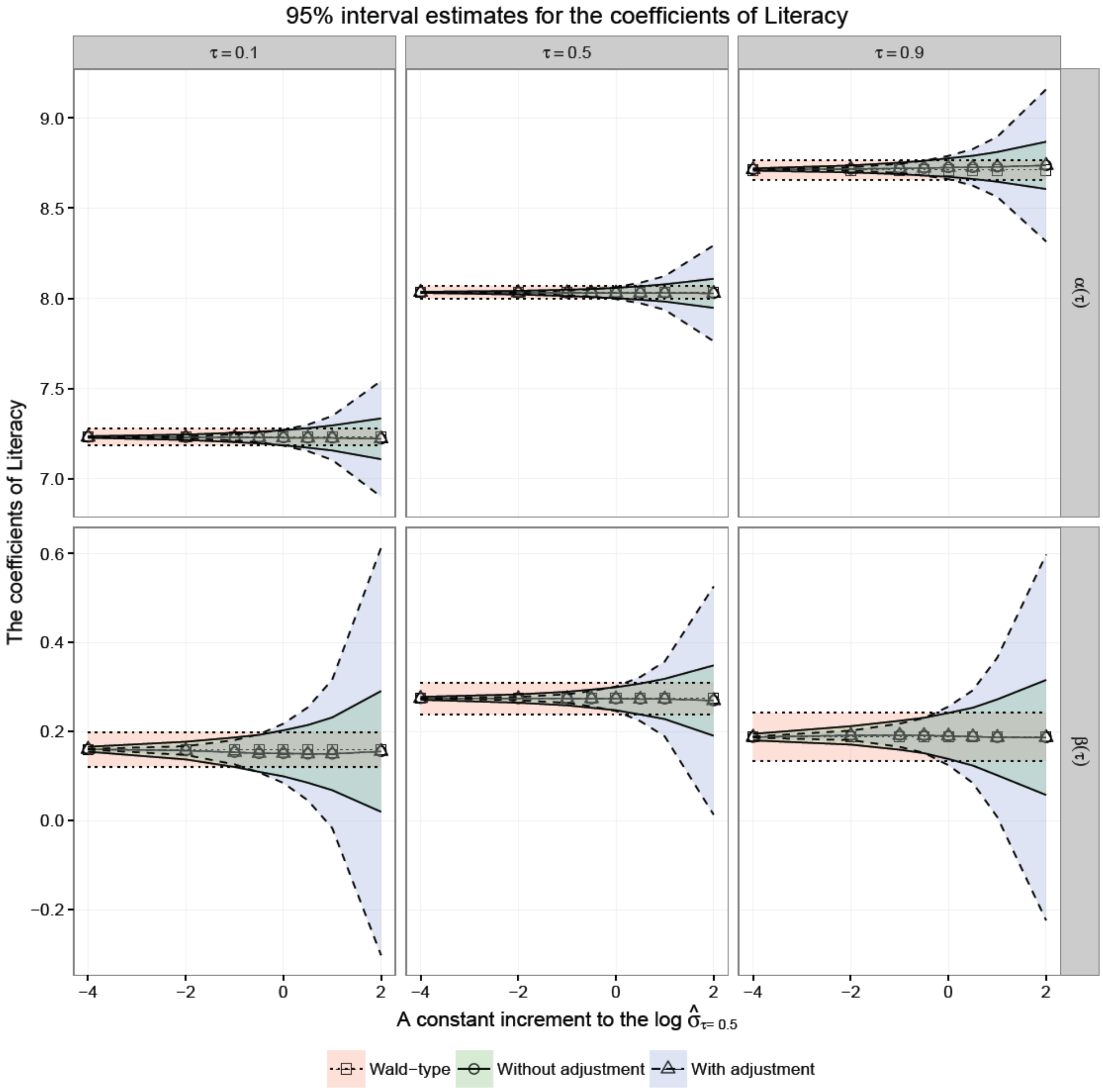

Figure 2.10: Sensitivity of the estimates of $\alpha(\tau)$ and $\beta(\tau)$ to the fixed scale parameter σ

Table 2.3: The point estimates and model-based standard error estimates (in parentheses) of $\alpha(\tau)$ and $\beta(\tau)$ at $\tau = 0.1, 0.5, 0.9$ as σ varies in Bayesian quantile regression with AL likelihood

Fixed σ	Estimator	$\tau = 0.1$		$\tau = 0.5$		$\tau = 0.9$	
		$\alpha(\tau)$	$\beta(\tau)$	$\alpha(\tau)$	$\beta(\tau)$	$\alpha(\tau)$	$\beta(\tau)$
$\log\hat{\sigma}_{\tau=0.5}$ - 4 = 0.004	BR_{nid}	7.233 (0.024)	0.160 (0.020)	8.034 (0.018)	0.274 (0.018)	8.711 (0.027)	0.188 (0.028)
	$\text{Bayes}_{\text{unadj}}$	7.232 (0.002)	0.160 (0.003)	8.034 (0.001)	0.274 (0.001)	8.714 (0.003)	0.187 (0.004)
	$\text{Bayes}_{\text{adj}}$	7.232 (0.000)	0.160 (0.000)	8.034 (0.000)	0.274 (0.000)	8.714 (0.001)	0.187 (0.001)
$\log\hat{\sigma}_{\tau=0.5}$ - 2 = 0.030	BR_{nid}	7.233 (0.024)	0.160 (0.020)	8.034 (0.018)	0.274 (0.018)	8.711 (0.027)	0.188 (0.028)
	$\text{Bayes}_{\text{unadj}}$	7.231 (0.007)	0.158 (0.010)	8.032 (0.005)	0.274 (0.005)	8.717 (0.010)	0.192 (0.011)
	$\text{Bayes}_{\text{adj}}$	7.231 (0.003)	0.158 (0.005)	8.032 (0.002)	0.274 (0.002)	8.717 (0.005)	0.192 (0.005)
$\log\hat{\sigma}_{\tau=0.5}$ + 0 = 0.229	BR_{nid}	7.233 (0.024)	0.160 (0.020)	8.034 (0.018)	0.274 (0.018)	8.711 (0.027)	0.188 (0.028)
	$\text{Bayes}_{\text{unadj}}$	7.228 (0.021)	0.150 (0.025)	8.029 (0.015)	0.273 (0.014)	8.724 (0.026)	0.191 (0.026)
	$\text{Bayes}_{\text{adj}}$	7.228 (0.022)	0.150 (0.032)	8.029 (0.017)	0.273 (0.015)	8.724 (0.033)	0.191 (0.033)
$\log\hat{\sigma}_{\tau=0.5}$ + 2 = 1.692	BR_{nid}	7.233 (0.024)	0.160 (0.020)	8.034 (0.018)	0.274 (0.018)	8.711 (0.027)	0.188 (0.028)
	$\text{Bayes}_{\text{unadj}}$	7.223 (0.060)	0.155 (0.070)	8.030 (0.042)	0.269 (0.041)	8.734 (0.067)	0.185 (0.066)
	$\text{Bayes}_{\text{adj}}$	7.223 (0.180)	0.155 (0.236)	8.030 (0.140)	0.269 (0.134)	8.734 (0.220)	0.185 (0.213)
$\log\hat{\sigma}_{\tau=0.5}$ + 4 = 12.503	BR_{nid}	7.233 (0.024)	0.160 (0.020)	8.034 (0.018)	0.274 (0.018)	8.711 (0.027)	0.188 (0.028)
	$\text{Bayes}_{\text{unadj}}$	7.181 (0.180)	0.190 (0.203)	8.020 (0.116)	0.260 (0.117)	8.791 (0.183)	0.186 (0.176)
	$\text{Bayes}_{\text{adj}}$	7.181 (1.562)	0.190 (1.989)	8.020 (1.081)	0.260 (1.110)	8.791 (1.620)	0.186 (1.494)
$\log\hat{\sigma}_{\tau=0.5}$ + 6 = 92.385	BR_{nid}	7.233 (0.024)	0.160 (0.020)	8.034 (0.018)	0.274 (0.018)	8.711 (0.027)	0.188 (0.028)
	$\text{Bayes}_{\text{unadj}}$	6.511 (0.937)	0.298 (0.806)	7.994 (0.361)	0.237 (0.405)	9.487 (1.005)	0.129 (0.784)
	$\text{Bayes}_{\text{adj}}$	6.511 (42.412)	0.298 (31.454)	7.994 (10.451)	0.237 (13.160)	9.487 (48.875)	0.129 (29.872)

Note: $\text{Bayes}_{\text{adj}}$ and $\text{Bayes}_{\text{unadj}}$ represent the Bayesian quantile regression based on an AL likelihood, with and without posterior variance adjustment; BR_{nid} is the frequentist quantile regression using Barrodale and Roberts's linear programming algorithm for l_1-regression with the Wald-type standard errors.

2.6 Conclusion

In this paper, we explored the sensitivity of posterior inference of Bayesian quantile regression to the value of the scale parameter of the AL likelihood. It is shown in the paper that not only the variance directly obtained from posterior distribution but also the adjusted posterior variance proposed by Yang et al. (2015) depend heavily upon the value of the scale parameter. In contradiction to Yang et al.'s (2015) argument that the adjusted posterior variance is asymptotically invariant to the value of σ, we found that the adjusted variance is even more sensitive to the scale parameter than the unadjusted one. While Yang et al.'s (2015) empirical evidence was based on only one data set, our results are based on both a real data analysis and comprehensive simulation studies that varied important factors such as data generating models. Our finding persisted for all data generating models considered in the simulation design.

It is important to note that the proposed posterior variance adjustment works only when σ is fixed at the maximum likelihood estimate of the scale parameter at the median ($\widehat{\sigma}_{\tau=0.5}$). Even though Yang et al. (2015) argues that the posterior variance adjustment leads to asymptotically valid posterior inference independent of the choice of σ, our finding suggests that fixing the scale parameter at $\widehat{\sigma}_{\tau=0.5}$ should be considered as a critical requirement to obtain valid posterior intervals. Although further work is needed to provide a theoretical proof, it seems that the posterior variance-covariance matrix does not approximate to $\sigma D_1^{-1}/n$ for finite n when σ is fixed. We found that the Bayesian approach performs reasonably with $\widehat{\sigma}_{\tau=0.5}$. This method can be a simple alternative to the Syring and Martin (2015)'s Gibbs posterior scaling algorithm that adaptively select the scale parameter to calibrate the corresponding Gibbs posterior variance.

We also found that point estimates can be biased both in conventional and Bayesian quantile regression at extreme quantiles. According to the asymptotic results for $\widehat{\boldsymbol{\beta}}(\tau)$ by Hao and Naiman (2007) the estimator is asymptotically unbiased. In practice, however, it has been often reported that the slope parameter estimates of linear quantile regression tend to be finite-sample biased at extreme quantiles (e.g. Cade et al., 2005; Sriram et al., 2013; Hausman et al., 2014). The Bayesian estimates tend to be more biased when the scale parameter is fixed at large values.

This might be due to a reduced density of the data in the tails. In quantile regression, it is known that away from the median the distributions of the estimated parameters become skewed and their dispersion is greater because of the data sparsity (Davino et al., 2014). When fitting quantile regression to finite samples, the skewed and dispersed distribution is likely to shift its center. Fortunately, the direction of the shift is predictable as previous research and this paper have shown. For example, the slope parameter estimates for the the lower quantiles tend to be biased upwards towards the median-regression coefficients, whereas the coefficients for the upper quantiles are biased downwards towards the median-regression. Researchers working with small to medium sample sizes should consider the direction and magnitude of possible biases at extreme quantiles.

Appendix

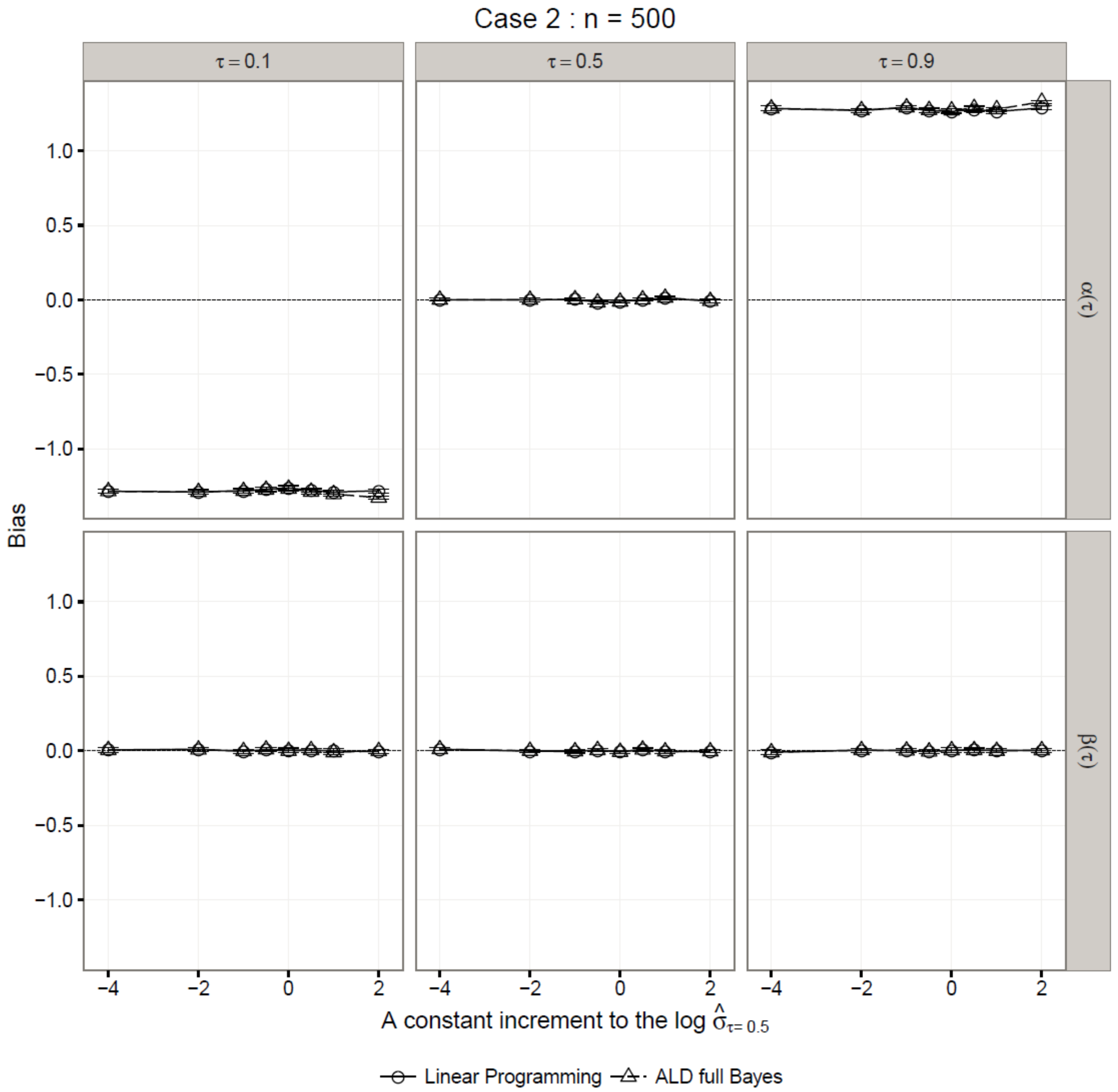

Figure 2.11: Sensitivity of bias to the fixed scale parameter σ in Case 2 with $n = 500$

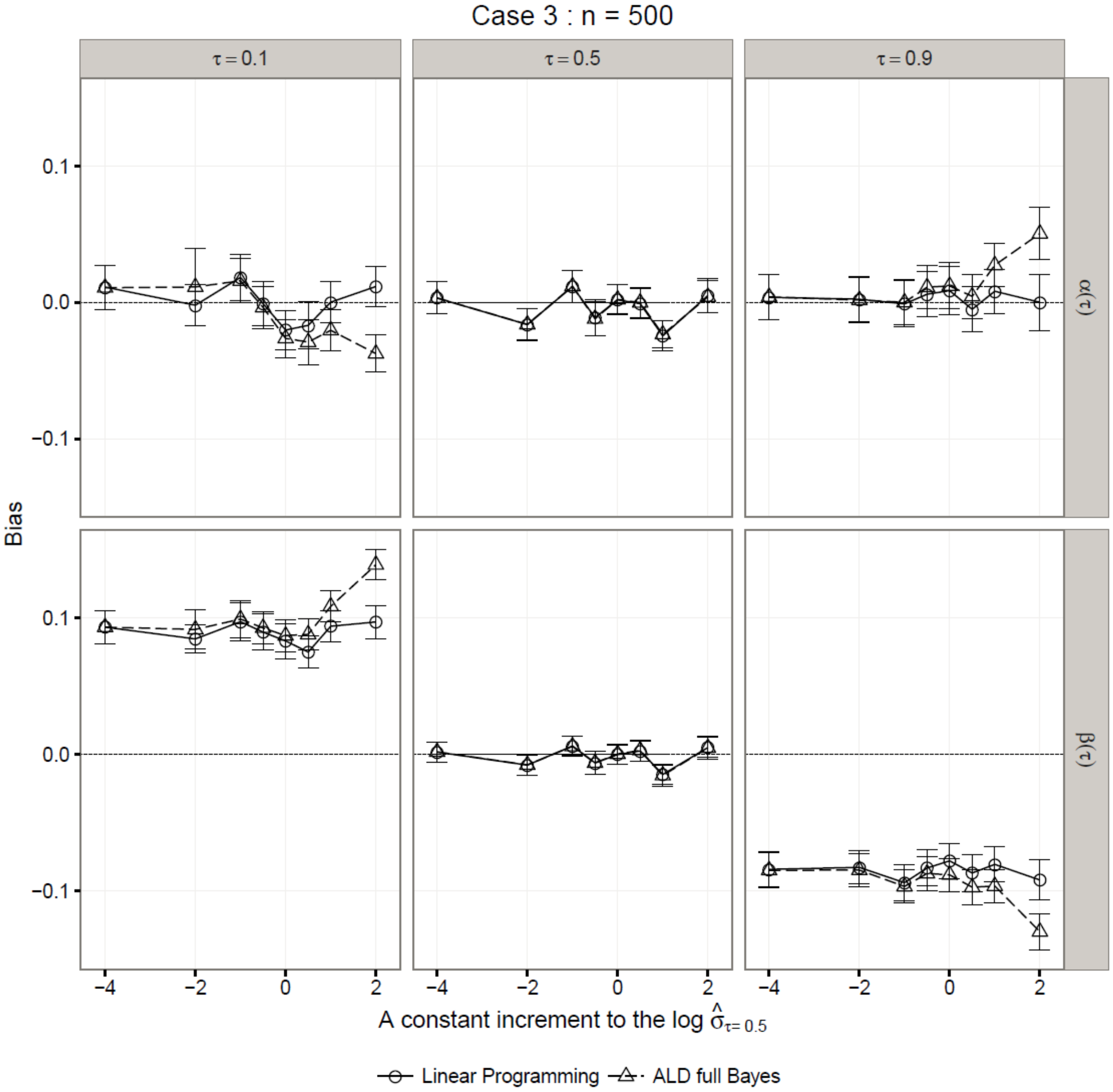

Figure 2.12: Sensitivity of bias to the fixed scale parameter σ in Case 3 with $n = 500$

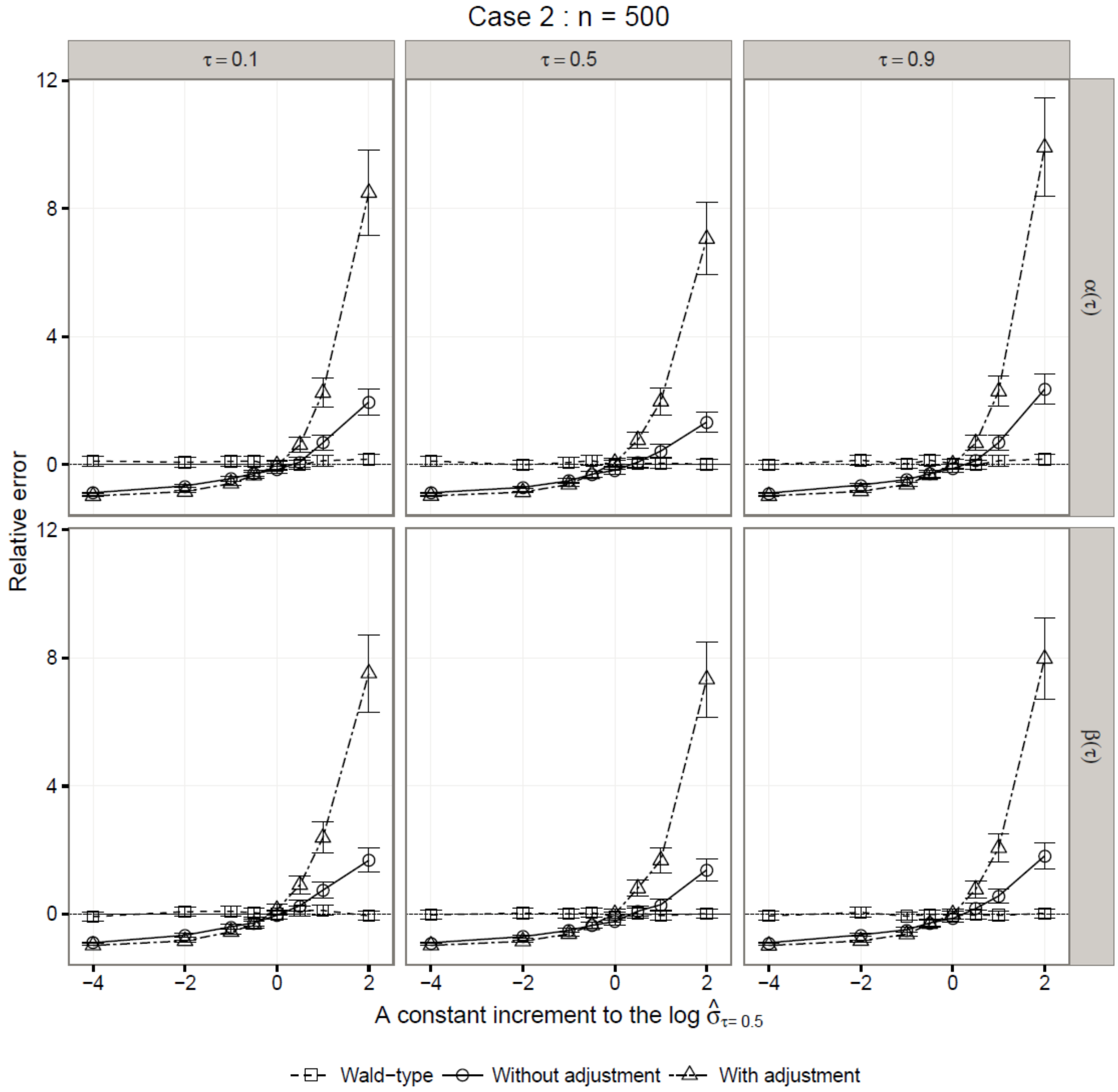

Figure 2.13: Sensitivity of the relative error of the model-based SE to the fixed scale parameter σ in Case 2 with $n = 500$

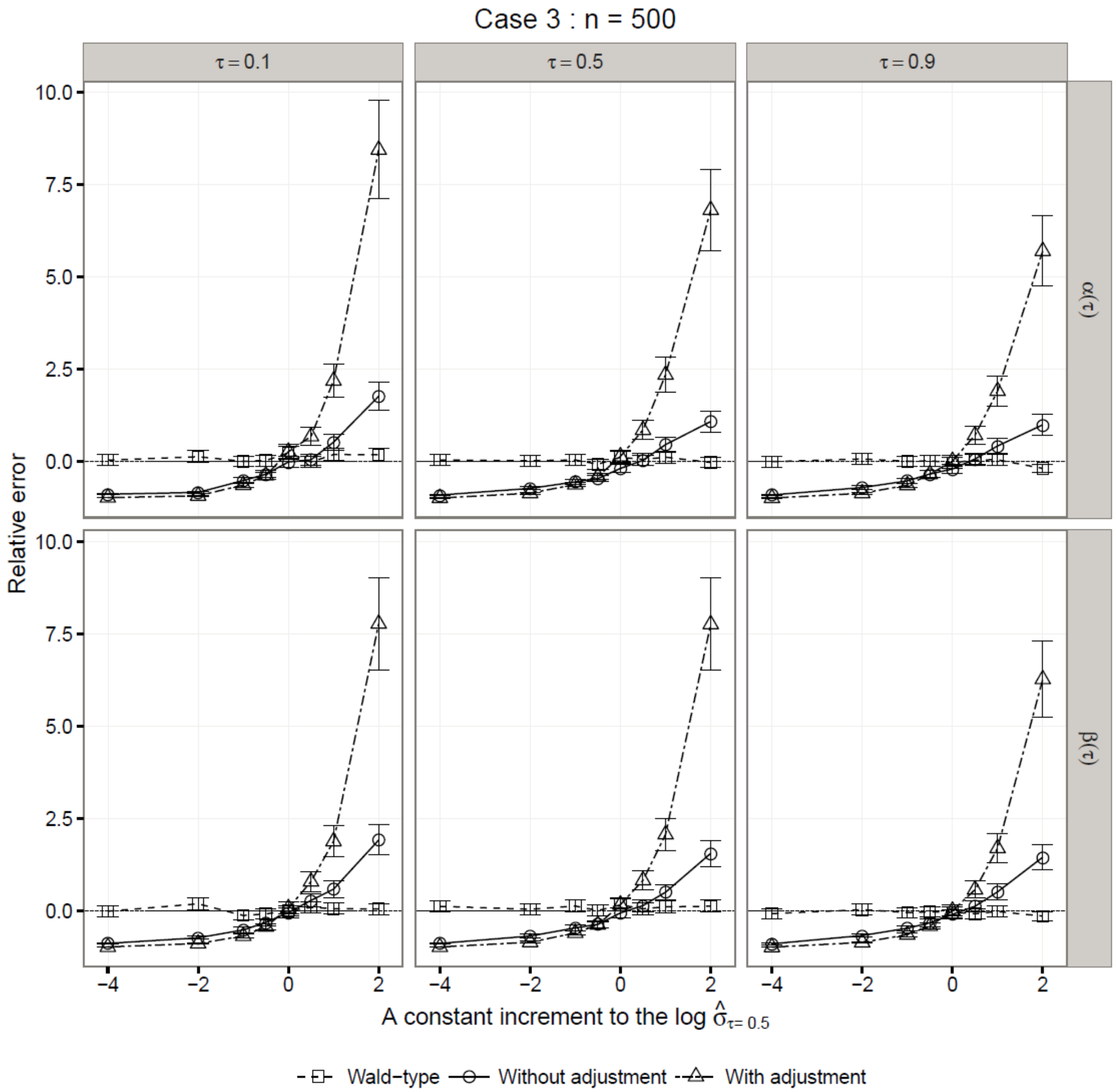

Figure 2.14: Sensitivity of the relative error of the model-based SE to the fixed scale parameter σ in Case 3 with $n = 500$

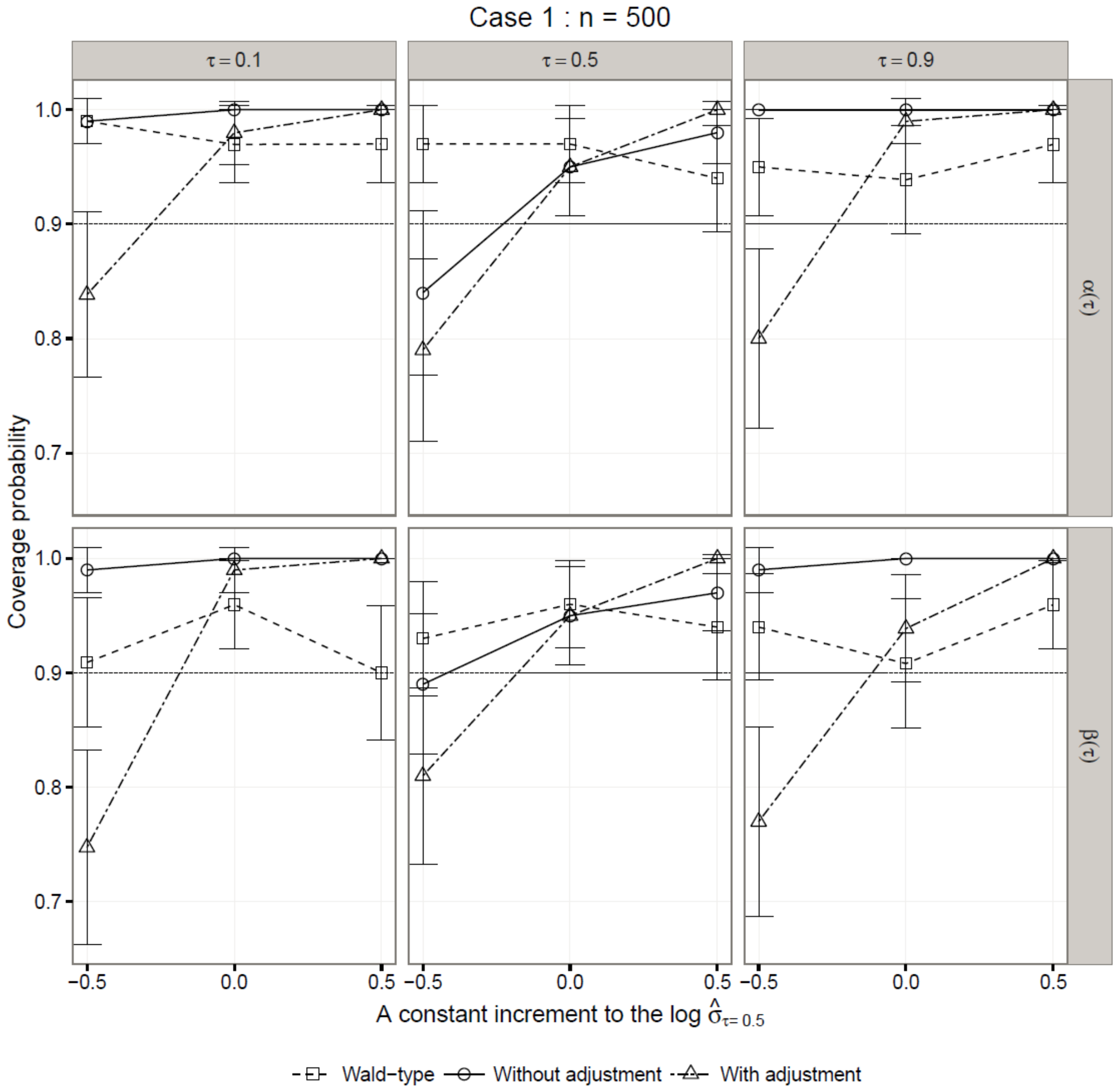

Figure 2.15: Sensitivity of the coverage of 95% confidence/credible intervals to the fixed scale parameter σ in Case 1 with $n = 500$

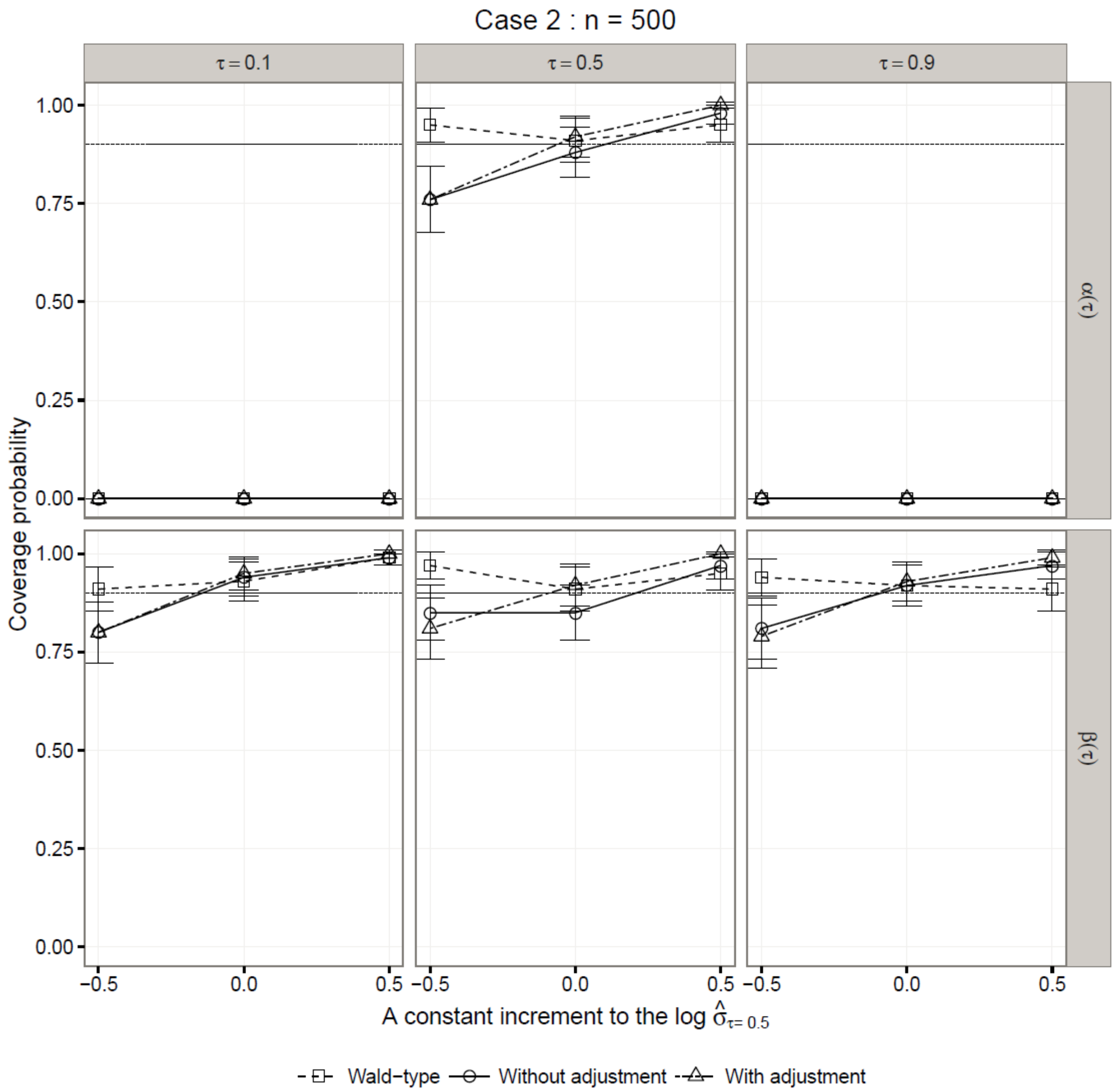

Figure 2.16: Sensitivity of the coverage of 95% confidence/credible intervals to the fixed scale parameter σ in Case 2 with $n = 500$

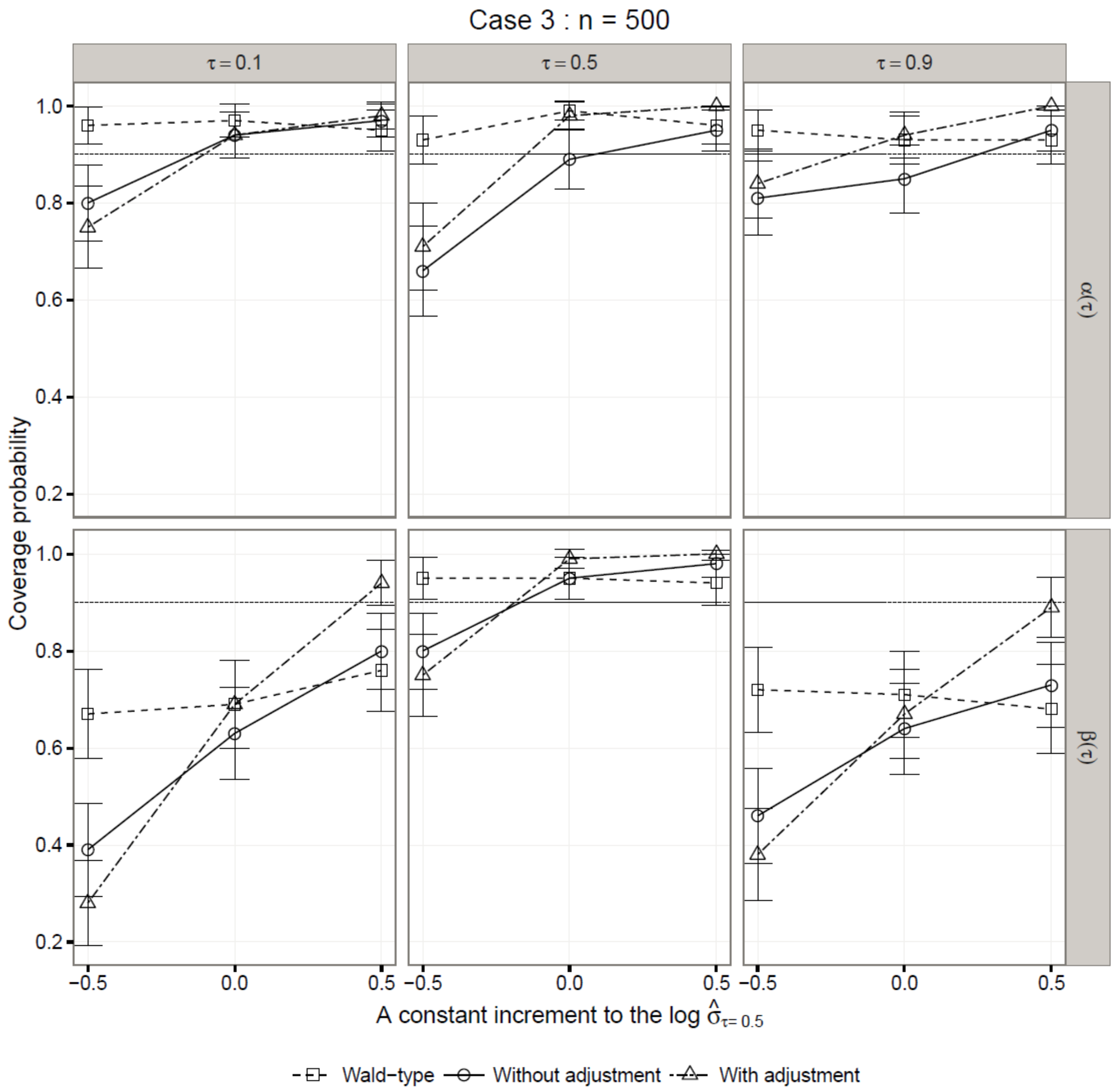

Figure 2.17: Sensitivity of the coverage of 95% confidence/credible intervals to the fixed scale parameter σ in Case 3 with $n = 500$

Chapter 3

Does finance Reform Move Teachers and School Organizations? California’s $23 Billion Equity Initiative[1]

3.1 Introduction

Fresh infusions of school funding work to raise student achievement, especially among disadvantaged pupils, according to findings emerging over the past quarter-century (Candelaria & Shores, 2019; Card & Krueger, 1992; Greenwald, Hedges, & Laine, 1996; Hyman, 2017; Lafortune, Rothstein, & Schanzenbach, 2018; Jackson, Johnson, & Persico, 2015). California’s contemporary reform offers one robust case in point.

Pressed by Gov. Jerry Brown, the California legislature approved in 2013 a weighted-student formula (Local Control Funding, LCF) that would boost annual spending by $23 billion by the fifth year, and send the lion’s share of new funding to districts serving large shares of English learners and pupils from poor or foster care families.[2] But little is known about how the effects of this ambitious reform, especially why it has failed to narrow achievement gaps, even in years when mean levels of learning climb (Johnson & Tanner, 2018; authors’ citation, in press).

We back-up to ask how infusions of new spending may shift the *attributes of teachers among* schools, and *organizational and curricular structures* that may variably direct inputs to certain students *within* schools. Despite calls in policy and scholarly circles for identifying not simply *whether* money matters, but *how* it alters teacher qualities or school-level mediators*,* the issue remains under-theorized and rarely studied (Hanushek & Woessmann, 2017).

This paper advances a conceptual model that first captures how this pair of intertwining mediators – teacher inputs and organizational practices – arrive to schools in varying levels and

[1] Sections 3.1-3.4 and 3.7 are co-authored with Prof. Bruce Fuller.

[2] California’s finance legislation and subsequent regulations refer to the reform as the *Local Control Funding Formula* or LCFF, an acronym that we shorten for brevity and to avoid confusion with *weighted-pupil formula* (WPF) in this paper.

mixes. Economists have long seen teachers as key inputs to schools, cast as a “production site”. But how new funding translates into inputs distributed among schools, and then deployed by principals within schools, is not well understood. We estimate how funding infusions may spur change in organizational practices, such as class size, teacher workloads, and the school’s relative emphasis on college-preparatory classes. Our theoretical framework also highlights how threats to the equitable distribution of inputs or practices can occur at three levels, as new dollars flow from state capitals: between districts, among schools within districts, and applied to differing students within schools.

The California reform operates through a weighted-pupil formula (WPF), granting wide discretion to local districts over how their new funds can be allocated, whether used for classroom instruction, rising health-care costs or pension liabilities. Pro-equity groups such as the American Civil Liberties Union (ACLU) have sued several districts that allegedly ignored the spirit of the LCF reform, allocating new dollars to schools or cost items disconnected from disadvantaged students (Kohli, 2016, 2019). This prompts a counter-hypothesis to government’s official theory-of-action. That is, the insufficient targeting of new funding on stronger teachers or improved classrooms and curricular structures may lead to null effects when it comes to raising mean achievement or narrowing disparities (e.g., United Way, 2018).

This paper closely follows recent methodological advances by Johnson and Tanner (2018; Jackson, Johnson, & Persico, 2015) to identify the exogenous portion of California’s LCF finance reform. Then, we estimate the quasi-experimental effect of the reform on changes in the qualities of teachers working inside schools, organizational features of schools, and curricular structuring, during the initial four years of LCF implementation, 2013-2017.

Building from Johnson and Tanner’s earlier study, we advance a theoretical framework and estimation strategy that (1) focuses on change in district-level *spending on students* and *instructional costs* (including teacher salaries), rather than relying on gross revenue changes, (2) taking into account multiple sources of possible endogeneity bias, (3) disaggregates elementary from high school effects, (4) examines how teacher inputs and organizational practices among schools may respond to varying levels of new dollars over time, and (5) employs quantile regression to illuminate how effects from finance reforms may differ among schools serving varying types of students.

We find that schools in districts that spent more on teachers and the instructional program over time, relative to their counter-factual level, at first hired more novice and substitute teachers and then continued to rely more on inexperienced teachers. The infusion of new LCF dollars did help high-poverty schools attract new, mostly white teachers with master’s degrees. This surge in hiring did foster a modest shrinkage of class size. Teacher workloads rose modestly in terms of the count of instructional periods assigned each day.

Additional effects may help to explain why California’s sharp rise in school spending has not discernibly moved achievement gaps. High schools increased the number of distinct courses, while reducing the share of courses qualifying as college-prep, especially in schools serving high shares of disadvantaged students. Growing shares of English learners attended classes taught by the rising count of novice teachers. Overall, we find that California's fresh infusion of dollars did spur the hiring of new teachers during the initial five years, but declining levels of teacher experience, along with observed organizational practices, worked against efforts to narrow disparities in student achievement.

We first review prior work that associates funding gains with achievement change, along with the few recent studies on school-level mediators. Second, we summarize what's known about decentralized and variably targeted finance reforms, some of which utilize weighted-pupil formulae (WPF). Third, we describe California's ambitious finance reform, spotlighting disappointing progress in narrowing achievement gaps, the stated goal of this massive reform.

3.2 Does finance reform alter school organizations?

3.2.1 The promise of progressive finance

Scholars continue to find achievement effects that stem from state-initiated gains in spending, efforts aimed at raising average pupil performance or narrowing disparities. Recent work, for instance, details how per pupil spending grew by over half in the poorest quintile of school districts nationwide (based on household income) between 1990 and 2012, yet by just under one-third in the most affluent fifth of all districts (Lafortune, Rothstein, & Schanzenbach, 2018). Encouragingly, students in the poorest districts displayed modest gains in achievement (about 0.10 SD over a decade, depending on grade level and subject) following discrete jumps in state spending. These researchers exploited the randomness of when states enacted finance reforms, allowing for quasi-experimental estimation.

Similarly, Jackson, Johnson, and Persico (2015) worked from randomly timed finance litigation among states, occurring in the 1970s forward, then estimated whether pupils achieved at higher levels over the post-reform period. Conducting this event study with instrumental variables, they estimated that a 10% boost in spending per pupil over 12 grades resulted in one-third of a year more schooling completed, along with a 7% bump in wages downstream. These benefits were greater for students from low-income families.

Narrower benefits of modest magnitudes have been estimated for California's LCF reform, first underway in the 2013-14 school year. Johnson and Tanner (2018) found that the exogenous portion of the LCF infusion of new dollars did predict significant increases in high school graduation rates and gains in eleventh-grade test scores in math and English language arts (ELA) over the initial four years. In math, students from poor families showed stronger gains than middle-class peers.

Overall, this line of work demonstrates that money *does* matter, at least when it comes to raising the learning curves of disadvantaged students. Yet, little empirical progress has occurred to discover *how* rising spending may alter the attributes of teachers deployed inside schools, or features of the school organization that may mediate achievement effects that resulting from new spending. Researchers have long tried to identify discrete inputs that display the greatest predictive validity (efficiency) vis-à-vis student achievement (e.g., Ferguson, 1991; Hanushek, 2016; Hedges, Picott, & Polanin, 2016; authors' citation). But this work has yet to be linked back up to state-level finance reforms, nor situated inside schools to learn how inputs are deployed by principals or teacher-leaders.

A handful of scholars are beginning to focus on mediating mechanisms or social processes that may stem from finance gains. Lafortune and SchÖnholzer (2018) found significant effects of school construction on achievement, as mostly disadvantaged children moved into new facilities in Los Angeles. Shifts in class size and teacher composition, differing between aging and new campuses, did not explain pupil-level effects, but the reduction in overcrowding did significantly mediate achievement gains, animated by a $19.5 billion boost in capital spending within the L.A. Unified School District.

Klopfer (2017) exploited the random incidence of state finance reforms since the 1970s to estimate change in indicators of school quality and instructional time. He found that finance infusions did not lead districts or schools to hire more teachers or better qualified instructional staff. But fresh funding did affect the length of the academic year, on average, allowing districts to add additional instructional days. This illustrates how funding infusions may animate organizational changes, which in turn may result in achievement gains.[3]

States may display less concern for efficiency under some finance reforms. In California, the governor and legislature placed priority on narrowing sticky achievement gaps by dramatically shifting new dollars to districts that serve larger concentration of disadvantaged students. In turn, Johnson and Tanner (2018) did find significant gains in teacher salaries and instructional spending, greater pension contributions by districts, and modest reductions in pupil-teacher ratios in the wake of LCF. Despite their pioneering work on district spending and school-level inputs, we know much less about the attributes of newly hired teachers, whether they are deployed to lift lower achieving students, and whether input infusions shift organizational and curricular structuring inside schools – possible mediators of how new spending lifts learning.

[3] He usually infers that districts and schools behave in ways to minimize labor costs: lengthen instructional time was about half as costly as hiring additional teachers.

3.2.2 Policy enthusiasm for decentralized weighted-pupil funding

Aiming to adequately fund public schools, while narrowing achievement gaps, policy activists and scholars increasingly embrace WPF strategies. Distributing state dollars based on a district's student composition (weighting certain students more heavily) recognizes the higher cost of lifting disadvantaged students over state proficiency hurdles. "Because not all students come to school with the same individual, family, or neighborhood advantages, some need more resources than others to meet a given achievement standard," argued the architects of California's reform (Bersin, Kirst, & Liu, 2008:5). This marks a shift away from *equal* to *greater* funding for disadvantaged pupils, as policy leaders struggle to narrow disparities.

WPF strategies also replace centrally regulated categorical-aid programs, limiting the state capital's role to the progressive distribution of funding, while decentering budget authority over programs and school-by-school allocations out to local districts (Augenblick, Myers, & Anderson, 1997; Odden & Picus, 2014). This proves politically appealing to district managers, who gain discretion, and unions leaders who welcome more money on the bargaining table. Then-Gov. Arnold Schwarzenegger pushed through legislation in 2009 to consolidate over 65 separate funding streams and programs. This precursor to the far-reaching LCF reform began to worry equity advocates, fearing diminished accountability over funds once targeted for poor children. After all, the original argument for categorical aid, going back to the Civil Rights Era, was that dollars would drift to schools in politically stronger neighborhoods, if left unregulated.

Evidence remains mixed on the extent to which WPF strategies affect the distribution of teachers, stronger school organizations, or curricular gains. Hawaii's statewide reform drove larger allocations to schools serving greater shares of poor students, but with few discernible effects for students (Levin et al., 2013). An ambitious WPF experiment in Prince George's County, Maryland triggered only a slight redistribution of resources to poor pupils. This district "unlocked" for principal discretion only certain teaching posts, while most remained centrally allocated (Malen, Dayhoff, Egan, & Croninger, 2015).[4] Two independent studies find that Los Angeles Unified progressively allocated more new dollars to high schools serving greater concentrations of poor students in the wake of new LCF funding. But per pupil spending climbed equally among elementary schools, whether situated in middle-class or poor neighborhoods (Partnership, 2018; United Way, 2018).

In California, Gov. Brown targeted a large amount of new dollars on big urban districts, along with other districts serving large shares of disadvantaged students. But his legislation placed no statutory strings on how districts could spend their new dollars. He resisted efforts by

[4] Miles and Roza (2006) similarly found that the devil lies in policy details, after studying WPF schemes in Cincinnati and Houston. The share of district budgets to which pupil weights are applied, niceties of the school-allocation formula, and highly institutionalized ways of distributing teaching posts worked to undercut redistributive effects.

the ACLU and equity advocates to require local boards to report on which schools benefit from rising state revenues. This leads to the pivotal policy question: do large infusions of new revenues – sent from a state capital out to districts in decentralized fashion – alter the attributes of teaching staff or the social organization of classrooms and curriculum in ways that improve mean achievement or narrow disparities in learning?

3.3 Progressive finance reform in california

> "*We are bringing government closer to the people, to the classroom where real decisions are made, and directing the money where the need and the challenge is greatest*." (Brown, 2013)

The Golden State's LCF reform swept aside scores of categorical aid programs in the summer of 2013, along with revenue limits on districts set in place by Proposition 13 back in 1978. The new weighted-pupil formula would contain three funding tiers. The *base grant* provides equal dollars per pupil in amounts set at about $6,900 per K-6 student, $7,200 per middle school student, and $8,300 per high school student, adjusted for inflation each year (California, 2013). *Supplemental grants* send to districts an amount equal to 20% of the base grant for each student from a poor family, designated EL, or child in foster care. *Concentration grants* further raise per-pupil distributions to districts by an amount equal to 50% of the base grant for each additional poor student after their representation surpasses 55% of total enrollment.[5] California regulations require that districts expand or improve services for the students generating the new *supplemental* and *concentration* grants in proportion to their share of district enrollment.[6]

3.3.1 Rising state spending

The state's resurging economy, along with a constitutional set-aside for K-12 spending, spurred a dramatic rebound from recession-era cuts, commensurate with Gov. Brown's approval of the LCF reform in the summer of 2013. Spending per pupil had declined by one-fifth through the Great Recession (2007-2011). Then, K-12 spending climbed by nearly $23 billion in yearly

[5] District leaders in Oakland and San Francisco had earlier experimented with WPF allocations among their constituent schools, awarding more dollars and discretion to schools serving larger shares of poor students. Effects on staffing patterns or pupil achievement remain mixed (Chambers, Shambaugh, Levin, Muraki, & Poland, 2008).

[6] Seeking transparency and compliance with this provision, the ACLU and pro-equity allies have won three cases to date where districts were found to divert new funding away from schools that served poor students (Fensterwald, 2015; Kholi, 2016).

outlays, rising to $61 billion in 2018-19, a remarkable boost relative to the pre-reform base year, 2012-13 (California, 2018). Per-pupil spending – at $11,645 – has reached an all-time high, although still far below states like New York and Massachusetts.[7]

The bulk of total state funding (88%) flows through the WPF mechanism, with additional dollars moving through categorical aid. The progressivity of the LCF allocation formula means that districts serving larger concentrations of poor students have received most of the new dollars. Districts with enrollments of less than 25% weighted students enrolled received just 5% more revenues per-pupil between, 2012-2019, compared with a one-third gain in revenues for districts with at least 80% of pupils falling into the weighted-pupil categories (EdSource, 2016).[8]

3.3.2 Little progress in narrowing achievement gaps

But despite this new spending, achievement gaps statewide have yet to budge. The percentage of Black and Latino students meeting the state's English-language arts (ELA) standard equaled 28% and 32% in 2015 (first year of California's shift to Smarter Balance testing), compared with 61% of White peers. This Black-White disparity had not moved through 2018, while the Latino-White disparity narrowed by 3 percentage points. In mathematics, no gap-closing was observed between Black and White pupils during the period. Since the second year of testing, overall student performance has been flat overall, with proficiency levels even falling among eleventh-graders, the only grade tested in high school (Department, 2018).

Importantly, our analysis reveals differing patterns between high and low-needs schools, based on differing shares of disadvantaged students (Figure 3.1). Panel A reports the percentage of high school pupils, split between English learners (ELs) and English-proficient students, who met or exceeded the proficiency standard in math. English-proficient pupils performed higher when attending low-poverty schools, those with less than half their enrollment disadvantaged. In general, gaps have failed to move. The pattern is similar when splitting students by race or economic status, except disparities widen for pupils attending high-needs schools.

[7] New York spent an estimated $22,433 per pupil in 2013-14; Massachusetts, $17,719 (Cornman, Zhou, Howell, & Young, 2018).

[8] The state's centrally set goals offer a mix of intentions, emphasizing implementation of Common Core State Standards; improving parent participation and school climate; widening "course access" and deepening student engagement, along with raising achievement and "other student outcomes" (California, 2013). The reform also hopes to widen civic participation in devising budgets, requiring a local accountability plan (Wolf & Sands, 2016).

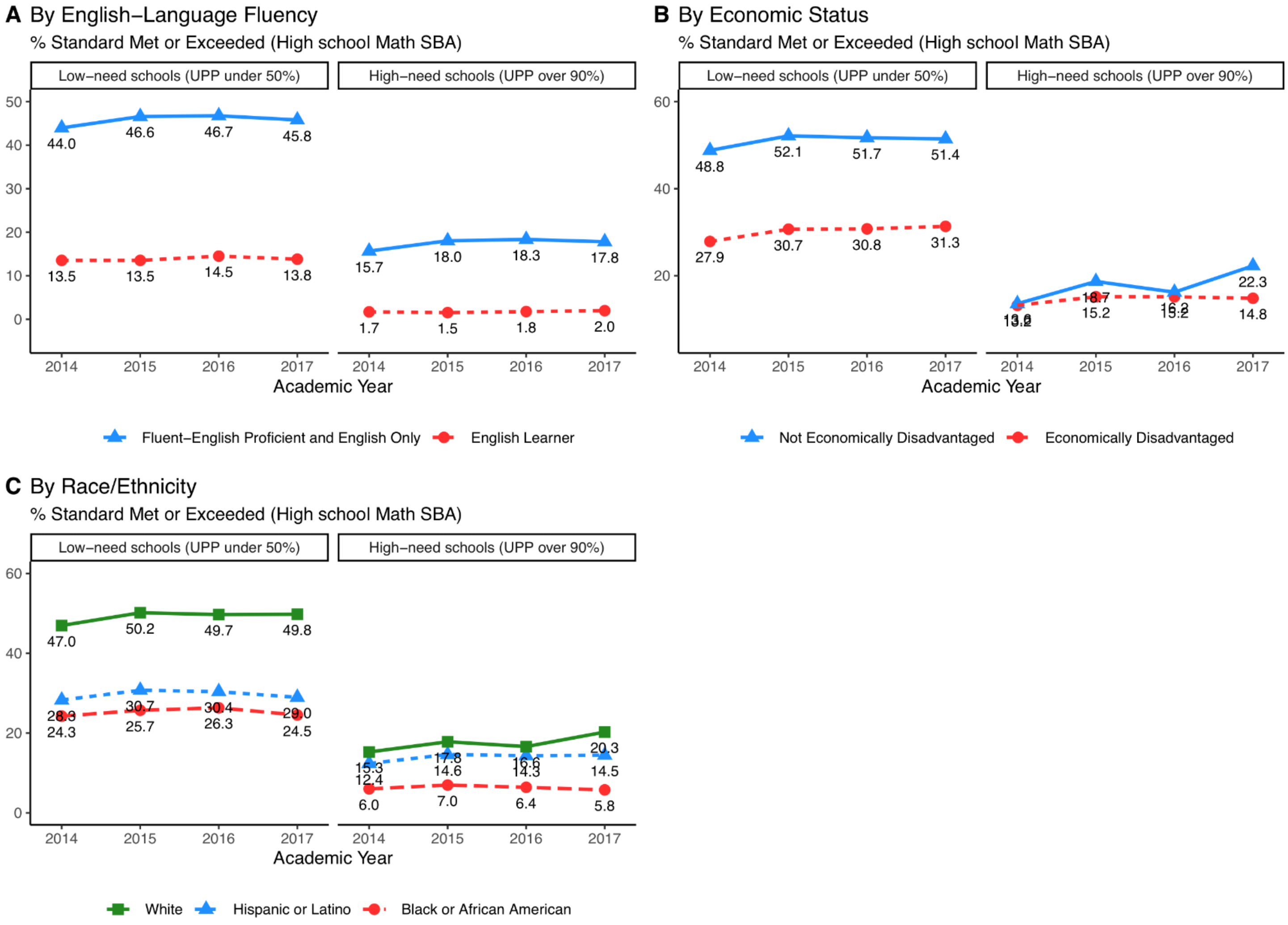

Figure 3.1: Change in high school mathematics Smarter Balanced state assessment (SBA) results by student subgroup.

Note: UPP represents the 5-year mean of the unduplicated percentage of disadvantaged students: free-or-reduced-price lunch eligible, English learners, and foster youth (2013-2017) as defined by California’s Local Control Funding reform.

Drawing on the same test-score data, Johnson and Tanner (2018) found that LCF-induced increases in district spending led to significant gains in eleventh graders' math achievement, particularly for disadvantaged students. Note in our panel B that the percentage meeting or exceeding standard in math rose from 27.9% to 31.3% for disadvantaged students in *low-poverty* high schools*;* in panel C Latino students in *high-needs schools* saw their average math performance move from 12.4% to 14.5% meeting or exceeding standard.

Note that Johnson and Tanner (2018) used "school and district-level averages that do not reflect inequality or changes across student groups within schools and districts (p. 27)," which may have missed heterogeneity of treatment effects among different types of schools. We observe from Figure 1 that school context matters, and likely yields differing effects from similar finance infusions. These descriptive patterns show how poor students attending schools in less-poor communities perform much better than peers attending schools surrounded by impoverished families. And achievement gaps between disadvantaged and non-disadvantaged pupils, and between white and black students, grew wider inside high-needs schools.

So, despite policy makers' aim of reducing disparities, California's massive infusion of new funding is somehow not getting to intended students, as districts distribute new dollars *among schools*, or principals *within schools* engage static or changing organizational practices and curricular structures. This problem animates the conceptual framework sketched in our analytic strategy and motivates the empirical analysis.

3.4 Analytic strategy

This section delineates a theoretical framework for that identifies pathways (or stages) through which fresh resources may prove effective in closing disparities in learning. Our analytic approach largely replicates and supplements the strategy employed by Johnson and Tanner (2018) to trace the effects of school finance reforms, and we closely follow their method of estimation. Yet, our analytic strategy moves from the disappointing fact that almost no progress has been made in narrowing achievement gaps in California.

First, we distinguish among three system levels at which threats to equitable distributions may occur, as states send new dollars down to local school districts (Figure 3.2). This distribution of inputs (or *resource patterns*) may be fairly or regressively distributed *among districts*. In California, "equitable" distribution recognizes the higher cost of lifting disadvantaged students over state proficiency hurdles. So, more dollars per pupil move to districts that host larger shares of disadvantaged students. Overall, our framework captures why finance infusions may not narrow achievement disparities – operating at three stages or organizational level: new resources flowing among districts; how districts distribute resources among schools; and how school leaders apply these fresh resources within organizational routines and practices.

In addition, we emphasize how new dollars *may not reach schools that host intended beneficiaries*. So, we must attend to the distribution of new dollars *among schools* within districts. Lafortune (2019) details how California's reform achieved progressive distribution of new funding to districts that host higher concentrations of poor students, but often fail to move new resources to constituent schools in proportion to their enrollment of disadvantaged pupils. Districts serving large concentrations of poor students, have been better able to lower class size, compared to low-needs districts, and allocate a larger share of their budget for teacher salaries and new hiring. Lafortune's findings mostly replicate Johnson's (2019:135-137) earlier results, now drawing on one additional year of implementation. But little work has traced which inputs in reality do attract new resources *inside schools* (recent exceptions discussed above, Klopfer, 2017; Lafortune & Schönholzer, 2018). So, we first examine whether schools in districts receiving greater funding (relative to their counter-factual level) employ differing mixes of teachers of terms of experience, employment and non-tenured status.

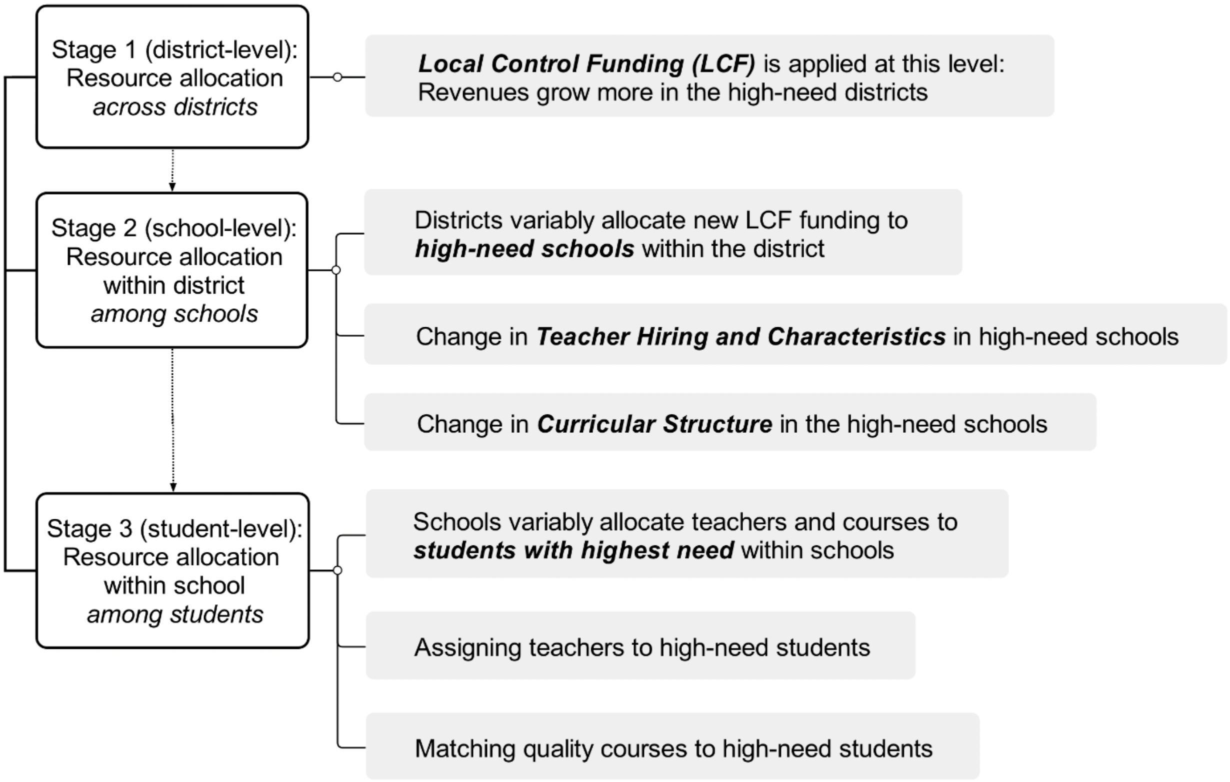
Stage 1 (district-level): Resource allocation *across districts*
Local Control Funding (LCF) is applied at this level: Revenues grow more in the high-need districts
Stage 2 (school-level): Resource allocation within district *among schools*
Districts variably allocate new LCF funding to ***high-need schools*** within the district
Change in ***Teacher Hiring and Characteristics*** in high-need schools
Change in ***Curricular Structure*** in the high-need schools
Stage 3 (student-level): Resource allocation within school *among students*
Schools variably allocate teachers and courses to ***students with highest need*** within schools
Assigning teachers to high-need students
Matching quality courses to high-need students

Figure 3.2: Three stages (or levels of social organization) in which inputs are distributed among districts, schools, and students.

We then turn to *features of the school organization*, including *curricular structuring*, that may further mediate the effects of new funding on the mean level and distribution of achievement among racial and social-class groups. In this way, our theoretical framework offers another conceptual advance: Estimating *which social practices inside schools are sensitive to funding infusions*, and which might mediate achievement effects. These factors are not only material inputs, they are mobilized or allocated within the rules and norms of human organizations, mediated by principals and other school leaders (Bryk et al., 2010; Simon, 1979; authors' citation). We address specific features of social organization of schools, including working conditions (pupil-teacher ratios, instructional periods assigned to teachers), along with decisions regarding the distribution of courses (e.g., curriculum that emphasizes college-prep courses or wide-ranging electives).

Finally, the effects of funding infusions on input allocations and organizational practices may vary among schools, based on the kinds of students being served. We saw above (Figure 1) how achievement levels and time-trends differ between high and low-needs schools. This suggests the utility of quantile regression to test for the differential effects of the district-level LCF policy at different points of the school-level outcome distribution within the same district.

Overall, we estimate the extent to which districts receiving larger funding increase altered teacher staffing and instructional inputs, and organizational practices, over the five years following the 2013 enactment of California's finance reform, speaking to these questions:

RQ1. To what extent did teacher characteristics, organizational practices, and curricular structure change among schools (descriptively) following 2013 enactment of California's finance reform through 2017?

RQ2. Does the exogenous portion of the finance reform help to explain change (from 2013 to post-reform years, 2014-2017) in teacher attributes, including the share of new teachers hired, reliance on long-term substitutes or less experienced teachers, and propensity to hold a graduate degree?

RQ3. Does the exogenous portion of the finance reform help to explain change in organizational practices and teachers' working conditions, including mean class size and the count of instructional periods assigned to teachers?

RQ4. Does the exogenous portion of the finance reform help to explain change in the curricular structure (as organizational practice) of high schools, including the prevalence of Advanced Placement or college-preparatory courses as shares of all courses offered, as distinguished from regular courses?

RQ5. Does the exogenous portion of the finance reform help to explain change access to experienced teachers or college-preparatory curriculum by English learners within schools?

RQ6. To what extent do estimated effects of the finance reform vary among schools within the same district, based on the socioeconomic features of students enrolled?

3.5 Method

3.5.1 Data

Core information comes from administrative data compiled by the California Department of Education (CDE), yielding measures of district-level revenues and spending, control variables, and the core outcomes measures related to teacher characteristics, organizational practices, and curricular structure. To exploit variation between schools, nested in all California school districts, we built school-by-year-level panel data sets, 2003-04 to 2016-17, for 6,867 traditional elementary and high schools (excluding charter schools) in 941 districts.

Data to identify the exogenous portion of the LCF reform. Local Control Funding "snapshot data" and district-level revenues and expenditures drawn from the standardized account code structure (SACS) data files are used to estimate the funding-formula-induced exogenous increases in district expenditures (CDE, 2018a). Data on district enrollment and the unduplicated count of students, used for supplemental and concentration grant calculations, are available from the LCF funding snapshot data. Monthly statements of general fund cash receipts and disbursements from the state's fiscal controller, including overall general fund spending, are used to construct the counterfactual trends in district per-pupil revenues, exploiting the exogeneity of the onset of the LCF reform.

To break down the SACS financial data into meaningful categories, we borrowed the definitions of expenditures utilized by Loeb et al. (2006), distinguishing student-related spending from other categories. This bin for student spending excludes district spending that's distant from classroom instruction, teacher salaries or student support services, such as debt service, capital outlays and facilities.

Data for student demographics and outcome measures. The California Longitudinal Pupil Achievement Data System (CALPADS) provides yearly data for teacher and staff demographics, pupils nested in courses (or elementary homerooms), and staff assignments by course, allowing us to construct a variety of outcome measures related to teacher characteristics and organizational features at the school level (CDE, 2018b). We linked CALPADS staff data files for each school year: (1) teacher demographics, experience, and credentials, (2) counts of FTE teachers, (3) teacher assignments to course and students, and (4) course enrollment data.

3.5.2 Measures

Using CALPADS data, we generated *school-level* outcome measures. Our school-by-year panel data (spanning 2003-2016) includes 79,688 school-by-year observations of 5,764 elementary schools and 14,972 observations of 1,103 high schools. Table 3.1 reports descriptive statistics for our three sets of outcome variables -- teacher characteristics, features or practices of school organizations, and curricular structure -- for elementary and high schools statewide. We include two control variables, enrollment counts and percentage of students eligible for free or reduced-price meals (FRPM). Data are split between the highest-poverty quintile (Q5) of schools and the lowest-poverty, economically best-off schools (Q1), based on 14-year averages of a school's share of FRPM students.

School enrollment and student attributes as controls. School enrollments and percentage of pupils, FRPM, were drawn each year from the CALPADS data. For elementary schools, we observe higher enrollments in high-poverty schools (mean enrollment, 627), than in low-poverty schools (558). The reverse is true for high schools.

Teacher characteristics. Variables for teacher attributes, aggregated to the school, are calculated from CALPADS staff data. This yields information on teacher counts by school and year and by whether the teachers were new to the district in a given year, the teachers' ethnicity, novice status (two or less years of experience), probationary or tenured status, and attainment of a master's or higher graduate degree.

Profiles of teachers differed between schools serving low or higher-income families. Only 39.2% of teachers in Q5 high schools, for instance, held master's degree or above, compared with 48.6% in Q1 high schools. About 8% of Q5 high school teachers were novices (less than two years experience), compared with 5% of teachers in Q1 schools. The ethnic composition of teachers differed sharply: under half the teachers in Q5 elementary schools were white (47.6%), compared with 84.5% of peers in Q1 elementaries. High schools showed similar differences.

Features of and practices within school organizations. Discrete organizational facets were calculated from CALPADS course-level data, including mean class size (again aggregated to the school level), and the mean number of class periods taught by each teacher, split for math and English language arts (ELA) classes, as two measures of working conditions. We see in Table 1 that mean class size is slightly smaller in high-poverty (Q5) high schools in ELA and math classes, than in low-poverty (Q1) schools. However, no significant differences in the mean number of class periods assigned to ELA or math teachers appear between Q1 and Q5 schools.

Curricular structure. CALPADS course-level data were used to generate the total number of courses and shares of classes designated as Advanced Placement (AP), and courses approved by the University of California for possible admissions. The latter are courses, falling into so-called 'A to G' categories, focused on core academic subjects, such as English, mathematics, and lab sciences. We refer to AP and A-G courses as *college preparatory*.

Table 3.1: Descriptive statistics for teacher characteristics, organizational features, and curricular structure by low- and high-poverty schools (Q1 and Q5), pooled data, 2003-2016

Variables	Elementary Schools			High Schools		
	Overall	Mean by Subgroup		Overall	Mean by Subgroup	
	Mean	Q1	Q5	Mean	Q1	Q5
Student characteristics						
Enrollment	548.5	557.6	627.2	1655.9	1893.1	1528.1
% Students eligible for FRPM	59.5	14.0	91.7	49.6	14.2	83.1
Teacher characteristics						
% Teachers newly hired in the district	7.4	7.1	6.9	10.5	9.3	11.0
Years of service in the district (school mean)	12.2	12.1	12.2	10.9	11.1	10.4
% Novice teachers (< 2 years of experience)	4.9	4.1	5.2	6.6	5.2	8.0
Teacher employment status:						
% Tenured teachers	78.5	80.7	76.9	73.2	77.2	70.2
% Long-term substitutes	5.6	7.0	4.6	5.5	7.0	5.2
% Probationary teachers	12.4	10.5	12.0	17.6	14.0	18.3
% White teachers	69.3	84.5	47.6	71.4	81.6	49.6
% Teachers holding a master's degree or above	35.3	40.7	33.3	40.8	48.6	39.2
School organization and working conditions						
School average class size (Homeroom)	23.2	23.9	22.9			
School average class size (ELA)				25.3	26.8	24.5
School average class size (Math)				26.4	27.9	26.3
Class periods assigned to teachers (ELA)				3.9	3.8	3.9
Class periods assigned to teachers (Math)				4.1	4.1	4.1
Curricular structure						
Total number of courses offered in the school	6.4	6.2	6.1	70.3	78.5	61.1
% Classes always approved as A-G (ELA)				7.2	9.0	6.1
% Classes always approved as A-G (Math)				55.3	66.9	49.8
% Classes cannot be approved as A-G (ELA)				6.7	4.4	9.4
% Classes cannot be approved as A-G (Math)				10.5	6.8	13.3
% AP classes (ELA)				5.8	7.5	4.6
% AP classes (Math)				5.1	8.5	3.4
Number of schools	5,764	1,145	1,144	1,103	220	219
Number of observations (school by year panel)	79,688	15,959	15,970	14,972	3,124	2,589

Note: Highest (lowest) poverty schools are those in the top (bottom) quintiles of school-level distributions of 14-year mean percentage of FRPM students (2003-2016), labeled as Q1 and Q5 respectively

The structure of the curriculum differs consistently between Q1 and Q5 schools. Campuses hosting higher-income families (low poverty) list 78.5 differing courses offered over the entire period, on average, compared with 61.1 in low-income counterparts.[9] In general, low-poverty high schools offer more college-preparatory courses than their high-poverty schools. The percentage of math classes that qualify for the A-G designation, for instance, is much lower in Q5 high schools (49.8%), than in low-poverty (Q1) high schools (66.5%).

English learners' access to experienced teachers and college-prep curriculum. To capture how students are differentially assigned to differing instructional resources *within schools* at the third stage, we linked the CALPADS course enrollment data to teacher characteristics via each teacher's unique ID, available from 2012 to 2017. We then constructed measures that capture ELs' access to instructional resources within schools.[10] First, we calculated a simple index that equals the mean percentage of ELs enrolled in classes taught by novice teachers (two or less years of experience), minus the mean percentage taught by experienced teachers (more than two years) within each school (the two adding to 100%). The same measure is constructed for classes taught by non-tenured and tenured teachers.

Table 3.2 reports descriptive statistics for these proportional differences. The English leaners' access to experienced or tenured teachers differs little in elementary homerooms. But the tendency for novice or non-tenured teachers to be assigned to classes with a larger percentage of EL students is greater in high-poverty high schools, compared with low-poverty schools. For example, the math classes in high-poverty high schools taught by novice teachers had 6.5% more ELs on average compared with the math classes taught by experienced teachers in the same school. The mean difference measure was calculated to be 2.5% for low-poverty schools.

We generated a similar measure to summarize ELs' access to A-G classes: the mean percentage of ELs enrolled in classes approved as A-G, minus the mean percentage of ELs in classes not approved as A-G within each school. The math classes approved as A-G in high poverty schools had 18.4% less EL students on average compared with the non-A-G math classes in the same school. This gap ranges lower in low-poverty schools.

These gap measures capture the disparate allocation of instructional resources, including

[9] The mean high school during the period offered between 8 and 12 different courses in each the following subjects: ELA, math, social studies, science. Eight foreign language courses were offered, along with 14 career and technical education courses, the fastest growing category post-2013. In addition, five tutorial and "instructional service" courses were offered on average.

[10] We focus on the within-school measures of English learners (ELs) in this study because the CALPAD course enrollment data includes only the student group among three LCF-defined high-need student groups: English learner, low-income student, and foster youth. Also, in Figure 1, achievement gaps were the widest between ELs and English-proficient students. Since we aim to uncover how fresh resources may prove ineffective in closing disparities in learning, examining within-school resource allocation to ELs is well-aligned with our original interests.

change in curricular structuring, within schools, which has not been explored in earlier studies. We use the similar school-level teacher quality indicators measuring teacher experience and education, similar to Johnson and Tanner (2018), while adding detailed aspects of teachers' employment status. And we examine the differential effects of LCF policy on different schools within a district, while Johnson and Tanner (2018) assumed commonly shared effects.

Table 3.2: Access of English learners to experienced teachers and college-prep (A-G) classes, pooled data, 2012-2017.

Variables		Overall	Mean by Subgroup	
		Mean	Q1	Q5
English learners' access to experienced teachers				
$\overline{\%EL}_{Novice} - \overline{\%EL}_{Experienced}$	Elementary homeroom classes	0.1	1.5	-1.1
	High school ELA classes	3.9	2.1	5.6
	High school Math classes	4.4	2.6	6.5
$\overline{\%EL}_{NonTenured} - \overline{\%EL}_{Tenured}$	Elementary homeroom classes	0.3	1.1	0.4
	High school ELA classes	3.9	2.3	5.9
	High school Math classes	4.1	2.3	8.1
English learners' access to rigorous curriculum path				
$\overline{\%EL}_{AG} - \overline{\%EL}_{NonAG}$	High school ELA classes	-31.8	-24.7	-38.4
	High school Math classes	-13.7	-10.2	-18.4

Note: (1) $\overline{\%EL}_{Novice} - \overline{\%EL}_{Experienced}$: The average percentage of English learners (ELs) in classes taught by the novice teachers minus the average percentage of ELs in classes taught by the experienced teachers (more than 2 years of experience) within the school. (2) $\overline{\%EL}_{NonTenured} - \overline{\%EL}_{Tenured}$: The average percentage of ELs in classes taught by the non-tenured teachers minus the average percentage of ELs in classes taught by the tenured teachers within the school. (3) $\overline{\%EL}_{AG} - \overline{\%EL}_{NonAG}$: The average percentage of ELs in classes approved as A-G minus the average percentage of ELs in classes not approved as A-G within the school.

Highest (lowest) poverty schools are those in the top (bottom) quintiles of school-level distributions of 14-year mean percentage of FRPM students (2012-2017) and are labeled as Q1 and Q5 respectively.

3.5.3 Empirical strategy

The key challenge in estimating effects that stem from progressively targeted finance, including WPF-style initiatives, is that school spending is an *endogenous* treatment. That is, school spending tends to be associated with unobserved time-varying or time-invariant school-level factors, either attributes of student or parental selection, forces that likely drive both school-level outcomes and funding gains. Thus, the primary challenge of causal identification is to isolate the arguably *exogenous* spending changes induced solely by the (LCF) finance reform. Two sources of exogeneity have been exploited to identify such reform-induced changes in per-pupil spending: the timing of reform events and the state's funding formula. [11] The novelty of Johnson and Tanner (2018)'s methodology lies in leveraging both sources of exogeneity by conducting an event study with a simulated instrumental variable (IV) approach. Their instrumental variables for estimating the effects of LCF on student outcomes are: (1) the number of school-age years a student was exposed to the LCF policy (*exposure*), and (2) the LCF reform-intended fully funded amount in district per-pupil spending from the state (*dosage* or *simulated IV*). While the former relies on the timing of LCF reform event being random or arbitrary, the latter exploits the availability of the formula for the *intended* allocation of funding along with the variables the formula was based on.

Let us review how Johnson and Tanner (2018) exploit these two sources of exogeneity to tease out only the LCF reform-induced funding increases, and use the exogenous variation to estimate potential effects of the spending increases on average student outcomes at the school level. Their design consists of three estimation steps: (1) prediction of the counterfactual district per-pupil revenue in the absence of the LCF, (2) estimation of the LCF-induced exogenous increases in district per-pupil expenditure, using dosage (or simulated IV), exposure, and the predicted revenues from the first step, and (3) estimation of the effect of the LCF-induced exogenous increases in district per-pupil expenditure on averaged school-level outcomes. Steps 2 and 3 correspond to the first and second stage of two-stage least squares IV estimation (2SLS-IV). This involves carving out a part of variation that is exogenous in the first stage, and then using only that part in to estimate causal impacts on an outcome in the second stage.

Step 1 is necessary for the causal identification when combining the 2SLS-IV with an event-study framework. The unobserved *time-invariant* district or school- level confounders and common statewide time trends in outcomes might be addressed by incorporating a variety of fixed effects and by instrumenting the district per-pupil spending with dosage and exposure. Still,

[11] This approach parallels two tandem lines of work. The first group of studies is based on the idea that states without reform events serve as a useful counterfactual for states that do have reform events, after accounting for fixed differences between the states and for common time effects. Assuming that the exact timing of events is as good as random, these studies employ an event-study framework and report a narrowing gap in Scholastic Achievement Tests (SAT) by parental education (Card & Payne, 2002), higher graduation rates for high poverty students (Candelaria & Shore, 2017), and a gradual reduction of family-income effects on national assessment scores (Lafortune, Rothstein, & Schanzenbach, 2018). The second group of studies leverage reform-induced variation in funding brought about by the funding formula within individual states. These studies are mostly based on instrumental variable type estimation. Guryan (2001) use sharp discontinuities in the state funding formula as exogenous instruments for district revenues and observe increases in fourth-grade math, science, and social studies test scores. Papke (2008) uses discontinuities in Michigan's funding formula to instrument for school expenditures and finds meaningful increases in the percentage of fourth-graders who cleared the state proficiency standard in mathematics.

unobserved *time-varying* confounders remain, namely the dynamic effect of structural economic conditions on district revenue and expenditures, which might confound the relationship between the LCF policy treatment and changes in school-level outcomes over time.

To account for these time-varying confounders, Johnson and Tanner modeled the predicted counterfactual evolution of K-12 revenues in the absence of the LCF. The rapid acceleration of state K-12 spending may stem both from the legislated framework under LCF and California's Proposition 98 funding guarantee, requiring that about 38% of the state budget go for public schools and community colleges. As California's economy continued to expand in the post-recession period, the district per-pupil revenues would have grown without the LCF policy enacted in 2013 due to Proposition 98 which reflects business cycle fluctuations. Johnson and Tanner directly predict this counterfactual trend based on prior funding and statewide California spending on non-K-12 expenditures.

Our analytic approach replicates Johnson and Tanner's three-step research design, acknowledging that their methodology is a generalized and sophisticated strategy for tracing the effects of California's LCF finance reform. We precisely replicated steps 1 and 2, with a single exception: we separate district total spending into *student* and *non-student* spending and use only the student spending for the predictor of focal interest. Student spending parallels the CDE's definition of Current Expense of Education per ADA and excludes spending on debt services, capital outlay and facilities, pre-K and adult programs, retiree benefits, and other non-agency and community services (consistent with Lafortune, 2019; Loeb, 2006; Bruno, 2018). Student-tied spending thus includes teacher salaries, instructional materials and supplies, special education, and pupil services. We estimate the effect of LCF-induced per-pupil *student spending* increases, which are more integral to pupil experience inside schools.

Our estimation method diverges from Johnson and Tanner at Step 3, since we do not estimate the effect of LCF-induced spending increases on birth-cohort-specific student outcomes, but instead estimate the effect on changes in teacher inputs and organizational practices applied to the entire school. Given our motivation of examining input levels and organizational practices in tandem, the analytical unit exposed to the LCF policy treatment is defined as a school, not a student, which leads to a diverging definition of the *exposure* variable. While Johnson and Tanner compares the change in (school-level average) student outcomes between exposed and unexposed "birth cohorts" from that district, we compare the change in school characteristics between exposed and unexposed "academic years" from that district. Both approaches share the common identification assumption that the timing of the LCF school finance reform is exogenous to changes in outcomes across different time points within districts. Since Johnson and Tanner use school-level averages of pupil achievement, the structure of the school-by-academic-year panel data constructed from our design is the same as the school-by-cohort panel data set used in their analysis, differing only in the time scale.

One methodological contribution appears in Step 3 where we use the recently developed multilevel instrumental variable quantile regression approach (Chetverikov, Larsen & Palmer, 2016) to estimate the *heterogenous* or *distributional* effects of the district-level LCF policy on school-level outcome distributions within districts. The common approach of estimating the location-shift of group-level averages of outcomes may mask important but subtle effects on the outcome distribution. In LCF reform studies, including Johnson and Tanner (2018), district-level increases in per-pupil spending may have little effect on the district-level average of school-level teacher quality measures but may still move the lower or higher quantiles of teacher quality distributions within a district. We allow each district to have a different treatment effect, and we estimate the treatment effect on district-level quantiles instead of on the district-level mean. We next provide the details of each step.

Step 1 – Predicting the counterfactual trends of district per-pupil revenue in the absence of LCF. The first step constructs the counterfactual trends in district per-pupil revenues. Following Johnson and Tanner, we construct two variables predict counterfactual district per-pupil revenue: (a) expenditures for total *state* operations, excluding education-related categories such as spending on state universities and colleges (State_t), and (b) the total *local* assistance expenditures outside of spending on K-12 schools, community colleges, and the state teacher retirement system (Local_t). These two variables were deflated by the consumer price index (CPI-U) to real 2016 dollars, divided by the state K-12 enrollment each year, and then converted to the natural log scale. The Step 1 prediction model is given by

$$\mathrm{logPPR}_{dt} = \alpha_{0,d} + \alpha_{1,d}\mathrm{logState}_t + \alpha_{2,d}\mathrm{logLocal}_t + \lambda_t + \epsilon_{dt}. \qquad (1)$$

where $\log PPR_{dt}$ is the natural log of district per-pupil revenue from the state for district d for year t; $\alpha_{0,d}$ is a district-specific intercept, λ_t is a year fixed effect; ϵ_{dt} is an error term. $\alpha_{1,d}$ and $\alpha_{2,d}$ represents the expected percentage change in the district per-pupil revenue for district d when State_t and Local_t increase by 1%, respectively. These coefficients encapsulate the district-specific sensitivity of revenues to changes in statewide expenditures.

The parameters of this model were estimated using the pre-LCF data (2003-2012) and then predictions were made for the post-LCF (2013-2017) years. Thus, the predicted log per-pupil revenue $\widehat{\mathrm{logPPR}}_{dt}$ in the post-LCF years can be viewed as an estimate of the counterfactual per-pupil revenue if LCF had not occurred (Johnson & Tanner, 2018). This reflects the dynamic effect of time-varying economic conditions on district revenues that might confound the relationship between LCF policy treatment and changes in school-level characteristics over time.

Figure 3.3 shows increases in observed per-pupil revenues during the pre-recession years (2003-06), along with dramatic reductions during the recession (2007-12), leading-up to LCF implementation. Predicted values in the post-LCF (2013-2017) years suggests that average per-pupil revenues would have increased without the LCF policy as the state expenditures recovered from the recession. The gap between observed revenue and its prediction in the post-LCF years can be regarded as an exogenous increase in per-pupil revenue due to the LCF reform. And recall how the LCF reform dramatically shifted the *distribution* of new funding to districts that serve larger shares of disadvantaged pupils.

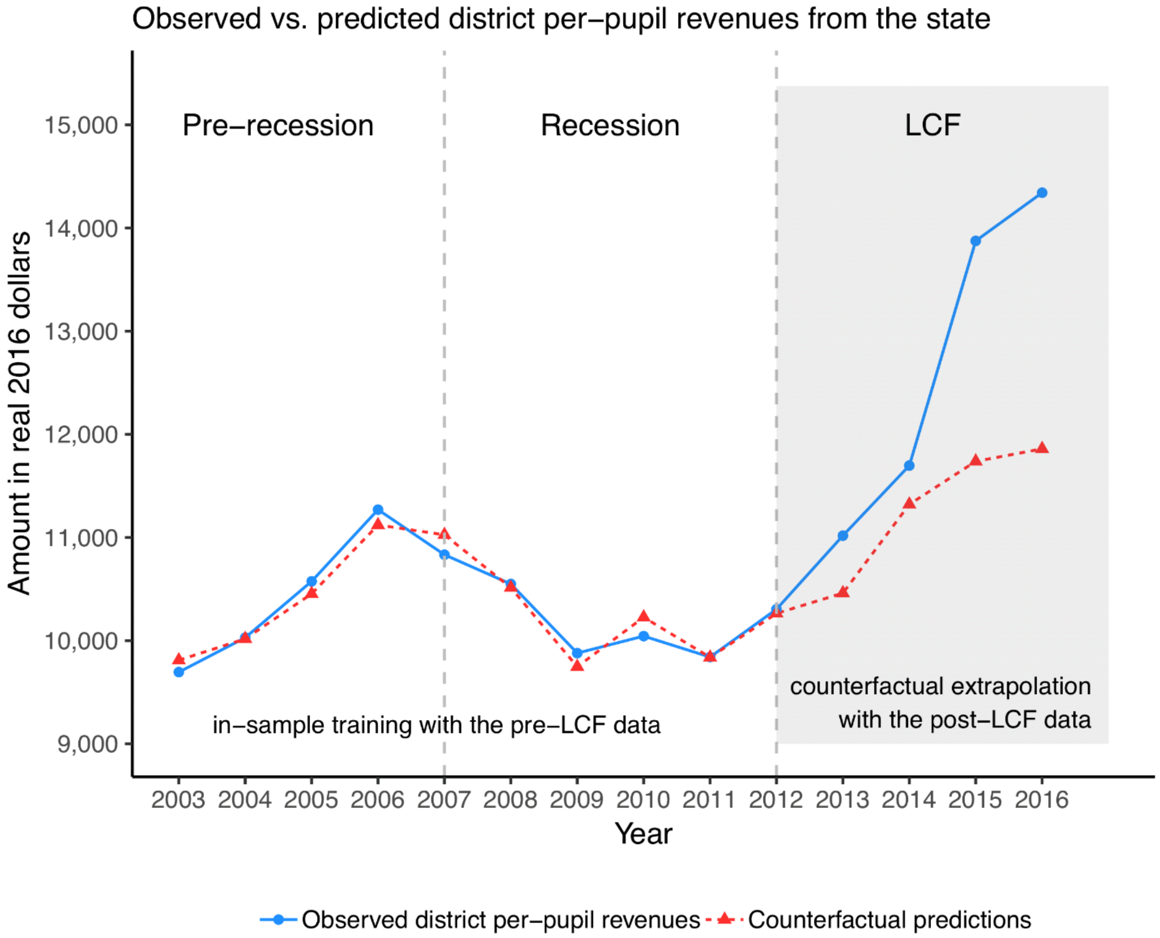

Figure 3.3: Observed and predicted (counterfactual) levels of California state revenues per pupil, 2003-2016. *Note*: Dollars are deflated into constant 2016 dollars.

Step 2 – Estimating LCF-induced exogenous increases in district per-pupil expenditure. To isolate exogenous changes in district per-pupil student spending that are unrelated to unobserved determinants of school-organizational features, we first construct two key variables containing the sources of exogeneity at this step: Dosage_d and Exposure_t. Dosage_d, *the simulated instrumental variable* or *dosage* for district d, is the LCF-intended amount of the supplemental and concentration grants in 2013, generated from the state funding formula. The CDE publishes the LCF Funding Snapshot every year to show key data elements and to summarize individual LCF target entitlement calculations for all school districts. LCF target entitlement refers to the target levels of LCF fully funded amount based on the following funding formula for district d:

$$\mathrm{Fund}_d = \mathrm{Base}_d + \{0.20 \times \mathrm{Base}_d \times \mathrm{UPP}_d\} + \{0.50 \times \mathrm{Base}_d \times \max[\mathrm{UPP}_d - 0.55,\ 0]\}. \quad (2)$$

Base_d is the base grant that depends on enrollment and varies only by grade level. UPP_d is the unduplicated percentage of disadvantaged students: those eligible for free or reduced-price lunch, with limited English proficiency, or in foster care. The supplemental grant is 20% of Base_d multiplied by the UPP_d. The concentration grant is an additional grant equal to 50% of Base_d for each district with UPP_d in excess of 55 percent multiplied by the district's UPP_d points above 55 percent. LCF funding snapshot data provides Base_d and UPP_d, which enables us to obtain the LCF-intended amount of funding, Fund_d.

Dosage_d is defined as $\mathrm{Fund}_d - \mathrm{Base}_d$ because it takes only the supplemental and concentration grants portion which is directly relevant to the overall level of district-level disadvantage. We use only the dosage for the first year of the reform (2013-14), Dosage_d^{2013}, to rule out any effects caused by district's incentive to classify more students as disadvantaged to obtain more funding. Next, $\mathrm{Exposure}_t$ represents the number of school years after the initial year of LCF reform for academic year t. $\mathrm{Exposure}_t$ varies from 0 (pre-LCF years from 2003, before 2013-14) to 4 (post-LCF year 2016-17) and reflects the exogenous timing of reform event.

Once the two key variables are defined, the first stage model of 2SLS-IV is estimated through the following fully nonparametric event-study model (following Jackson, Johnson, & Persico, 2015; Johnson & Tanner, 2018):

$$\mathrm{logPPE}_{dt} = \sum_{z=1}^{10}\sum_{p=0}^{4}\left(I_{\mathrm{Exposure}_t=p} \times I_{\mathrm{DoseDecile}_d^{2013}=z}\right) \cdot \alpha_{p,z} + \gamma_1 \widehat{\mathrm{logPPR}}_{dt} + \mu_d + \lambda_t + v_{dt}. \quad (3)$$

The endogenous treatment variable of interest, logPPE_{dt}, is the natural log of per-pupil student spending for district d for year t. District fixed effects μ_d and year fixed effects λ_t are included to account for general underlying differences across districts and years and to exploit only variation within district-by-year cells. $\widehat{\mathrm{logPPR}}_{dt}$ is the predicted natural log of the counterfactual per-pupil revenue for district d for year t estimated from step 1 to capture time-varying confounders.

To estimate a flexible version of the first-stage equation of 2SLS-IV rather than assuming linear slopes of $\mathrm{Exposure}_t$ and Dosage_d^{2013}, we convert each original variable to a series of indicator variables, $I_{\mathrm{Exposure}_{t=p}}$and $I_{\mathrm{DoseDecile}_{d=z}^{2013}}$, respectively. $I_{\mathrm{Exposure}_t=p}$ equals 1 if

Exposure_t equals p which varies from 0 to 4, and 0 otherwise. $I_{\text{DoseDecile}_{d=z}^{2013}}$ is an indicator variable that takes a value of 1 if the decile of Dosage_d^{2013} ($\text{DoseDecile}_{d=z}^{2013}$) equals z and 0 otherwise. Thus, the coefficients for the two-way interactions of $I_{\text{Exposure}_{t=p}}$ and $I_{\text{DoseDecile}_{d=z}^{2013}}$, $\alpha_{p,z}$, summarize the LCF-reform-induced exogenous increases in per-pupil student spending in districts with dosage decile z after p years from the reform. $\widehat{\log\text{PPE}}_{dt}$ is then the natural log of per-pupil student spending for district d for year t instrumented by the two sources of exogeneity, the timing of reform event and funding formula.

Step 3 – Estimating the effect of LCF-induced increases in funding on the within-district distribution of teacher and school-organization outcomes. Once we carve out the part of per-pupil spending variation that is exogenous in the first stage of 2SLS-IV at Step 2, we specify the second-stage outcome model at Step 3 to estimate the effect of the LCF-induced exogenous increases on school outcomes. The structure of the outcome model follows a difference-in-difference (DiD) estimation approach. Instead of having treatment and control groups, the LCF-induced exogenous spending increases serve as the "the amount of treatment" or "dosage". We can obtain the DiD estimate by comparing the pre-to-post intervention change in school-level outcomes between *high dosage districts* and *low dosage districts*.

In such DiD estimation, A conventional linear model can be fit to estimate the treatment effect on the district-level *mean* of the school-level outcome variable. This common approach of estimating the location-shift of district-level averages of outcomes, however, may miss heterogeneity of treatment effects among different types of schools within districts. The multilevel instrumental variable (IV) quantile regression approach (Chetverikov, Larsen, & Palmer, 2016) allows us to estimate such within-district heterogeneity. The goal of this model is to examine whether the LCF-induced district-level spending increases have differential effects at different points (quantiles) of the school-level outcome distribution within the same district, while controlling for unobservable district-level confounders. [12] Within the DiD framework, the conditional quantile, $Q_{Y_{sdt}}(\tau)$, at quantile level τ (e.g., $\tau = 0.2, 0.5, 0.8$) of the outcome variable, Y_{sdt}, for school s within district d in year t is modeled as

$$\text{Level-1:} \quad Q_{Y_{sdt}}(\tau) = \alpha_{dt}(\tau) + \beta_1(\tau)\text{Enroll}_{sdt} + \beta_2(\tau)\text{FRPM}_{sdt}, \quad \tau \in (0, 1) \qquad (3)$$

[12] The causal identification relies more on the validity of instruments, that is, how to validly tease out LCF-induced exogenous spending increases at Step 2, rather than the statistical modeling choices on the grouped outcome distribution at Step 3. But we make one additional assumption at Step 3 for the interpretation of estimates for a specific quantile. An impact estimated from the quantile regression basically measure the impact on a particular statistic (a quantile of interest), not necessarily on a specific school. To interpret the quantile regression estimates as differential effects of an intervention for individual schools, it is required to assume that school's rank within districts is preserved before and after the intervention, or stated differently, the LCF-induced spending increases do not substantially reorder schools in terms of school-level teacher or course characteristics (Schochet, Puma, & Deke, 2014).

$$\text{Level-2: } \alpha_{dt}(\tau) = \gamma_0(\tau) + \gamma_1(\tau) I_{\text{Exposure}_t > 0} + \gamma_2(\tau) \log \widehat{\text{PPE}_{dt}} +$$

$$\delta^{\text{DiD}}(\tau) \cdot \left(I_{\text{Exposure}_t > 0} \times \log \widehat{\text{PPE}_{dt}}\right) + \gamma_3(\tau) \log \widehat{\text{PPR}_{dt}} + u_{dt}(\tau). \quad (4)$$

Here, the cluster is defined as a district-by-year cell. Hence Equation 3 is referred to as the Level-1 or within-cluster model, and Equation 4 is referred to as the Level-2 or between-cluster model. For a fixed quantile level τ, The key term of this model is the varying intercept $\alpha_{dt}(\tau)$, interpretable as the district-by-year-specific conditional quantile of the school-level outcome after adjusting for differences between clusters in the level of the two school-level confounders: total enrollment (Enroll_{sdt}) and percentage of FRPM students (FRPM_{sdt}). Each district-by-year cell (cluster) has one value of $\alpha_{dt}(\tau)$.

$\alpha_{dt}(\tau)$ is treated as the outcome variable in the Level-2 between-cluster model to estimate the effect of the key treatment variable $\log \widehat{\text{PPE}_{dt}}$ where $u_{dt}(\tau)$ represents unobserved factors at the district-year level which can affect the τth quantile of Y_{sdt}. We are primarily interested in estimating the difference-in-difference parameter $\delta^{\text{DiD}}(\tau)$. $I_{\text{Exposure}_t > 0}$ equals 1 if Exposure_t is larger than 0, which indicates that the academic year is after the LCF. $\text{log}\widehat{\text{PPE}_{dt}}$ represents the predicted or instrumented natural log of district per-pupil student spending for district d in year t. Thus, the coefficient of their interaction term, $\delta^{\text{DiD}}(\tau)$, measures how much the change in $\alpha_{dt}(\tau)$ between exposed and unexposed academic years from the same district tends to be larger for those districts that experienced more LCF-induced increases in per-pupil spending across exposed and unexposed years, after controlling for the effect of time-varying confounder $\log \widehat{\text{PPR}_{dt}}$.[13] A positive value of $\delta^{\text{DiD}}(\tau = 0.2)$, for example, would indicate that the LCF-induced increase in per-pupil spending boosted the lower tails of the within-district distribution of the school-level outcomes after the reform.

We then extend this DiD model at level 2 to the fully nonparametric event-study model to account for the LCF's multiyear phase-in timeline to incrementally close the gap between new target level of funding and actual funding over years (following Johnson & Tanner, 2018):

$$\text{Level-2: } \alpha_{dt}(\tau) = \lambda_t(\tau) + \{\gamma_2(\tau) + \delta_t(\tau)\} \cdot \log \widehat{\text{PPE}_{dt}} + \gamma_3(\tau) \log \widehat{\text{PPR}_{dt}} + u_{dt}(\tau) \quad (5)$$

where $\delta_0(\tau) = 0$. Equation 5 includes the term $\gamma_2(\tau) \log \widehat{\text{PPE}_{dt}}$ and the constraint $\delta_0(\tau) = 0$ so that the parameter $\delta_t(\tau)$ represents the difference in the effect of $\log \widehat{\text{PPE}_{dt}}$ on $\alpha_{dt}(\tau)$ between reference year 2012-13 ($t = 0$, the year prior to June 2013 enactment of LCF or the pre-reform base year) and t years after (or before) the reference year after controlling for the effect

[13] The difference in logs can be used to approximate proportionate change. Because the treatment variable is in *natural logs*, the $\delta^{\text{DiD}}(\tau)/100$ represent the absolute change in $\alpha_{dt}(\tau)$ when $\widehat{PPE_{dt}}$ increases by 1%. This can be interpreted as the semi-elasticity of $\alpha_{dt}(\tau)$ with respect to $\widehat{PPE_{dt}}$. The effect of 10% increases in $\widehat{PPE_{dt}}$ can be approximated by $\delta_t(\tau)/10$. We would not want to use this nonlinear approximation for much larger percentage changes in $\widehat{PPE_{dt}}$ such as 50%.

of the time-varying confounder $\log \widehat{\mathrm{PPR}_{dt}}$. We present the nonparametric event-study estimates $\hat{\delta}_t(\tau)$ using the event study plots in Figure 3.7-3.9. The event study plots show the varying effects of the LCF-induced spending increases across the post-LCF phase-in period while the $\hat{\delta}^{\mathrm{DiD}}(\tau)$ presents only an average effect over the same period. The event study plots also allow us to visually assess the credibility of our research design by checking pre-reform estimates. Null effect estimates of LCF-induced spending increases in pre-reform years imply that the estimated LCF-induced increases in spending are not relevant to the changes of outcomes before the reform.

3.6 Findings

3.6.1 Time trends

Let us first examine trends in our three sets of outcomes – teacher characteristics, organizational features and curricular structure at the school level – before and after 2013 implementation of California's LCF finance reform.

Teacher characteristics. We first display time trends for teacher characteristics, aggregated to elementary schools statewide, then split between low-and high-poverty campuses (Figure 3.4). Overall, we observe that the hiring of novice teachers, new to the district, picked-up as new funding came to districts and schools. But these staffing profiles, in general, return to pre-recession patterns. The post-2013 surge in hiring also led to a modest spike in the hiring of long-term substitutes (not shown), but this trend line settled back down, as districts and schools hired novice, yet credentialed teachers. High-poverty schools tended to rely more on novice teachers in the post-reform period, compared with low-poverty schools.

One exception appears in panel D, where high-poverty schools have come to employ a higher share of tenured teachers, relative to heavier reliance on probationary (non-tenured) staff prior to the recession. This may be the result of leveling student enrollment statewide, and stabilizing teacher staffs. These time trends are similar for high schools (available from authors).

School organization, working conditions and practices. We observe more notable shifts, post-reform and relative to the pre-recession periods, when it comes to organizational features and curricular structure, the latter at the high school level. To illustrate the shrinkage of class size, we display this outcome for elementary homerooms, where students spend most of their time (Figure 3.5). Mean class size declined after the LCF reform, from 25.6 to 24.0 in low-poverty homerooms, and 24.4 to 23.3 in high-poverty schools, a modest improvement (panel A). These class sizes remain higher than in the pre-recession period, 2003-2008.

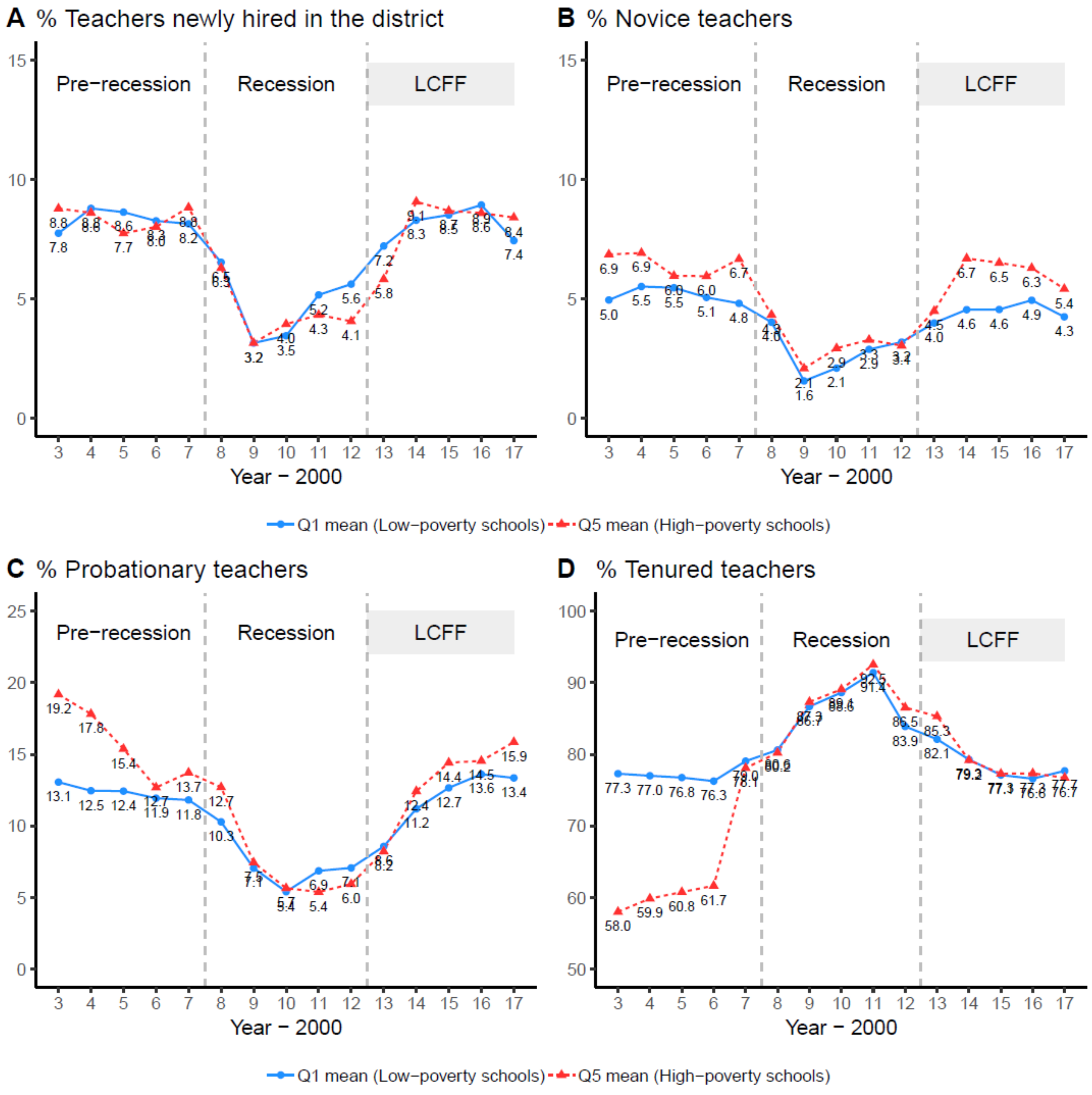

Figure 3.4: Time trends in selected teacher characteristics for low- and high-poverty elementary schools (Q1 and Q5, respectively), 2003-2017.

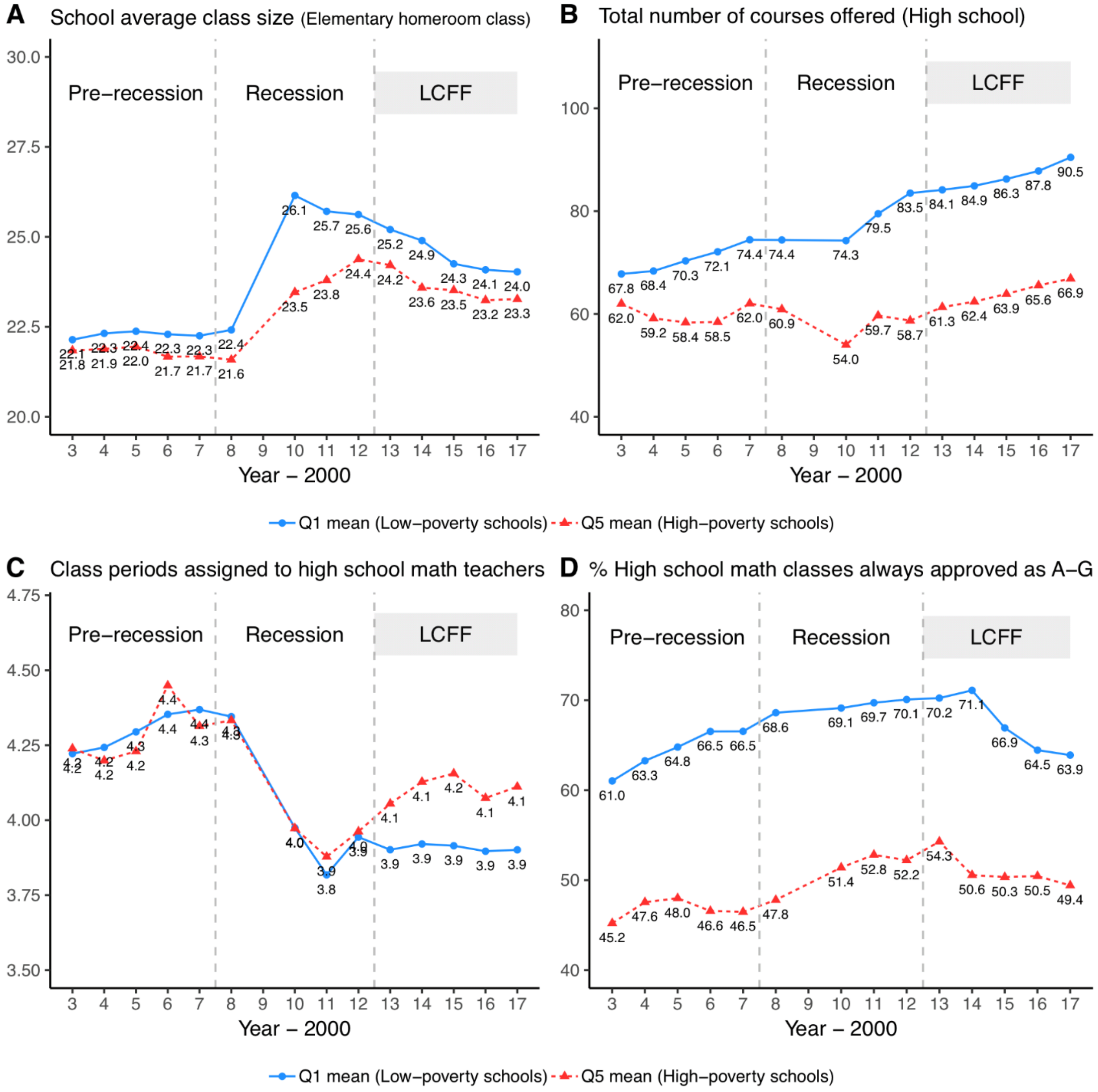

Figure 3.5: Time trends for selected organizational features and curricular structure for low- and high-poverty elementary and high schools (Q1 and Q5, respectively), 2003-2017.

Turning to high schools, the mean count of instructional periods assigned to teachers moved up slightly in high-poverty high schools (panel C) from a low of 3.9 during the recession to a high of 4.2, post-reform. No change in this count is observed for teachers in low-poverty high schools, and their mean count is considerably lower post-reform, compared with pre-recession workload levels.

Curricular structure. We see in panel B in Figure 3.5 that the mean number of course offered steadily increased, although counts remain lower in high-poverty schools. The curricular structure is becoming more differentiated in low-poverty schools, even relative to the pre-recession period. This may be a reaction to the earlier curricular narrowing under No Child Left Behind, along with observed growth in tutorial and special instructional periods appearing in the CDE course data.

Panel D suggests that high schools have modestly moved away from college-prep curriculum since the post-2013 infusion of new funding. As the listing of courses has grown among high schools statewide, a diminishing percentage are approved as A-G college-prep in character. And this structuring differs starkly between low- and high-poverty high schools. The share of courses deemed A-G declined from 70.2% to 63.9% during the five years since LCF implementation in low-poverty high schools. This diminishing trend fell from 54.3% to 49.4% in high-poverty high schools.

English learners' access to experienced teachers and college-prep curriculum. Access to experienced teachers and A-G courses by EL students also slipped during the post-2013 period (Figure 3.6). Panel A shows that the proportional representation of EL pupils in classes taught by non-tenured (probationary) versus tenured teachers varies little in low-poverty high schools. But EL representation in classes taught by non-tenured staff increased modestly during the post-reform period in high-poverty schools. Ideally, this disparity would narrow, if the reform aimed to reduce disparities in student achievement.

Panel B displays a similar pattern, with lower representation of EL students attending A-G courses in the year prior to the LCF reform, a gap that grows modestly post-2013. The practical magnitudes of these differences are not great. But they are consistently moving in the opposite direction of organizational or curricular shifts that would likely help narrow achievement gaps, at least for English-learners.

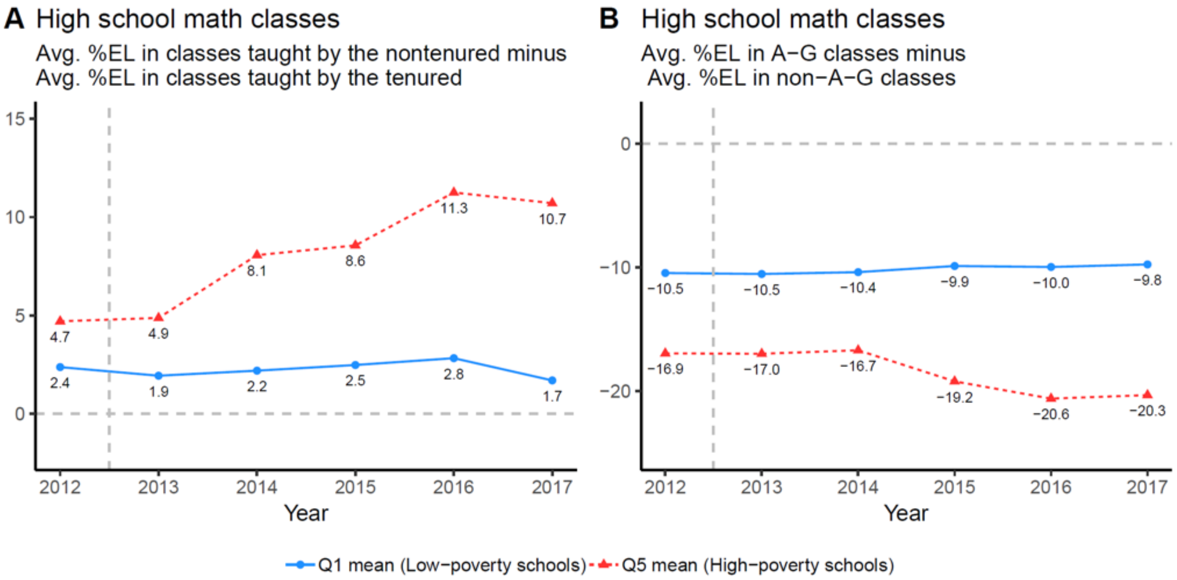

Figure 3.6: Time trends in access to non-tenured teachers and college-prep (A-G) classes by English learners, 2003-2017.

3.6.2 Did LCF move mediators? Results from the quantile estimation

Next we report difference-in-difference (DiD) estimates of LCF treatment effects for each set of school-level outcomes: teacher characteristics, school organization and teachers working conditions, curricular structure, and ELs' access to instructional resources within each school. Table 3.3 provides a summary of difference-in-difference estimates, $\delta^{\text{DiD}}(\tau)$, from the model equation 4 . Estimates shown in Table 3.3 indicates the effect of an 1% LCF-induced increase in the district per-pupil spending on the change in the district-specific conditional quantile of school-level outcomes after the event.

We observe that LCF-induced increases in per-pupil spending result in significant increases in the percentages of teachers who were newly hired to their district, as well as share of teacher workforces comprised of novice teachers. These findings are consistent with the significant reduction in average years of service in one's district. An LCF-induced 10% increase in district per-pupil expenditures during the post-LCF years resulted in a 0.98 percentage point gain in the increases in the share of newly hired teacher in elementary schools that previously had fewer share of new teachers during the pre-LCF years ($\tau = 0.2$). The effect size diminishes as the quantile level increases to $\tau = 0.8$ in elementary and high schools, which implies that schools that had more shares of new teachers in pre-LCF years hired fewer new teachers post-LCF.

Table 3.3: Summary of the estimated difference-in-difference estimates of the effects of LCF-induced spending increases on school-level teacher characteristics, organizational features, and curricular structure

Outcome variables	Elementary Schools			High Schools		
	τ(0.2)	τ(0.5)	τ(0.8)	τ(0.2)	τ(0.5)	τ(0.8)
Teacher characteristics						
% Teachers newly hired in the district	9.8^{***}	8.6^{***}	7.1^{***}	9.6^{***}	10.3^{***}	7.9^{***}
Years of service in the district (school mean)	-2.2^{***}	-2.7^{***}	-3.1^{***}	-0.3	-1.2^{*}	-1.6^{**}
% Novice teachers	3.7^{**}	2.9^{**}	2.7^{**}	5.0^{**}	5.4^{***}	4.7^{**}
% Tenured teachers	-6.6^{**}	-7.7^{**}	-7.8^{**}	-7.0^{*}	-11.2^{**}	-14.2^{**}
% Long-term substitutes/ temporary employee	1.5^{*}	1.9^{**}	2.3^{**}	0.3^{*}	0.4^{*}	0.4^{*}
% Probationary teachers	11.9^{***}	10.4^{***}	9.6^{***}	9.8^{**}	9.7^{**}	4.7^{*}
% White teachers	8.0^{***}	5.5^{*}	3.0	14.4^{***}	8.1^{*}	0.6
% Teachers holding a master's degree or above	4.0^{**}	2.2	1.0	8.5^{*}	8.7^{*}	5.9
School organization and working condition						
School average class size (Homeroom)	-3.3^{***}	-3.5^{***}	-3.8^{***}			
School average class size (ELA)				-2.3^{**}	-2.0^{**}	-2.1^{**}
School average class size (Math)				-2.1^{**}	-1.7^{*}	-2.0^{**}
Class periods assigned to teachers (ELA)				-0.0	-0.1	-0.1
Class periods assigned to teachers (Math)				0.5^{**}	0.2	0.1
Curricular structure						
Total number of courses offered in the school	0.4	0.1	1.3	3.9^{**}	4.5^{**}	6.9^{***}
% Classes always approved as A-G (ELA)				-3.6^{***}	-3.6^{***}	-5.7^{***}
% Classes always approved as A-G (Math)				-3.0	-8.6^{*}	-11.6^{**}
% Classes cannot be approved as A-G (ELA)				0.2	0.7	-0.6
% Classes cannot be approved as A-G (Math)				5.1^{**}	5.0^{**}	2.6
% AP classes (ELA)				-2.6^{***}	-2.7^{***}	-3.4^{***}
% AP classes (Math)				-2.2^{**}	-2.6^{***}	-3.2^{***}
English learners' access to experienced teachers and rigorous curriculum path						
$\overline{\%EL_{Novice}} - \overline{\%EL_{Experienced}}$ (Homeroom)	1.5	0.4	-2.0			
$\overline{\%EL_{Novice}} - \overline{\%EL_{Experienced}}$ (ELA)				7.6	4.6	-1.3
$\overline{\%EL_{Novice}} - \overline{\%EL_{Experienced}}$ (Math)				-1.9	2.8	8.8^{*}
$\overline{\%EL_{NonTenured}} - \overline{\%EL_{Tenured}}$(Homeroom)	-0.6	1.6	2.2			
$\overline{\%EL_{NonTenured}} - \overline{\%EL_{Tenured}}$ (ELA)				0.0	2.1	7.2
$\overline{\%EL_{NonTenured}} - \overline{\%EL_{Tenured}}$ (Math)				-1.9	1.0	3.7
$\overline{\%EL_{AG}} - \overline{\%EL_{NonAG}}$ (ELA)				9.3^{*}	2.8	-3.0
$\overline{\%EL_{AG}} - \overline{\%EL_{NonAG}}$ (Math)				3.1	-1.2	-4.5

Notes: ***p ≤ .01, **.01 < p ≤ .05, *.05 < p ≤ .10. The reported estimates are relevant to the difference-in-difference parameter, $\delta^{\text{DiD}}(\tau)$, from the model equation 4. Standard errors available from the authors.

Nonparametric-event-study estimates presented in Figure 3.7 (along with 95% confidence intervals) show the varying effects of the LCF-induced funding increase across the post-LCF implementation period. Here we only present results from the 0.5 quantile level (median), since patterns were quite similar in the other quantiles. Note that none of the effect estimates in pre-LCF years are statistically significant at the 5% level. This means that the estimated LCF-induced increases in funding are not relevant to the changes of outcomes before the reform and therefore lends support to our analytic strategy's ability to isolate the effect of the LCF-induced exogenous funding increases.

Effects of the LCF reform on the share of teachers newly hired were felt immediately in the first year of funding increases, and then peaked at the fourth year of reform (panel A). This pattern is mirrored by the reduction in average years of service in the district (panel B). And LCF-induced 10% increase in funding leads to the 0.6% points increase in the share of teaching staff, novices, after four years of exposure to LCF (panel C). Newly hired teachers, stemming from LCF-induced funding increases, often included non-tenured staff, such as novices.

Districts relied on hiring long-term substitutes in the first and second years of LCF, and then more on probationary teachers in subsequent years. The magnitude of effects from LCF-induced funding was the largest for the share of teachers with probationary status: 10% increases in funding lead to 1.75 percentage point increase in the share of probationary teachers after four years of LCF exposures (panel E). To put these magnitudes in context, per pupil spending grew by up to 40% during the post-LCF period in urban districts with high shares of disadvantaged students (California Analyst, 2018). So, a portion of these effects hold practical significance in the staffing and organizational structure of schools.

Figure 3.8 displays heterogeneous effects of LCF on the within-district distributions of school-level teacher composition. Panel A shows that an LCF-induced 10% increase in per-pupil spending results in about 1.12 percentage point and 1.05 point gain in the share of school staff made-up of white teachers at the lower tail of its distribution ($\tau = 0.2$) in the third and fourth years of LCF, respectively, while no significant effect was found for the higher quantile ($\tau = 0.8$).

A similar pattern was observed for share of teachers holding master's degree or above (panel A). This means that the exogenous funding increases lifted the percentage of school-level teaching staff that were, white or holding a master's degree or above, especially for schools that previously employed small proportions of such teachers in the reference year. Since the schools with lower shares of white or master's holders tend to serve higher-poverty students, as seen in Table 1, it's fair to conclude that the infusion of new LCF dollars helped high-poverty schools to attract more white or better highly-educated teachers, mitigating any disproportionate sorting of teachers between schools.

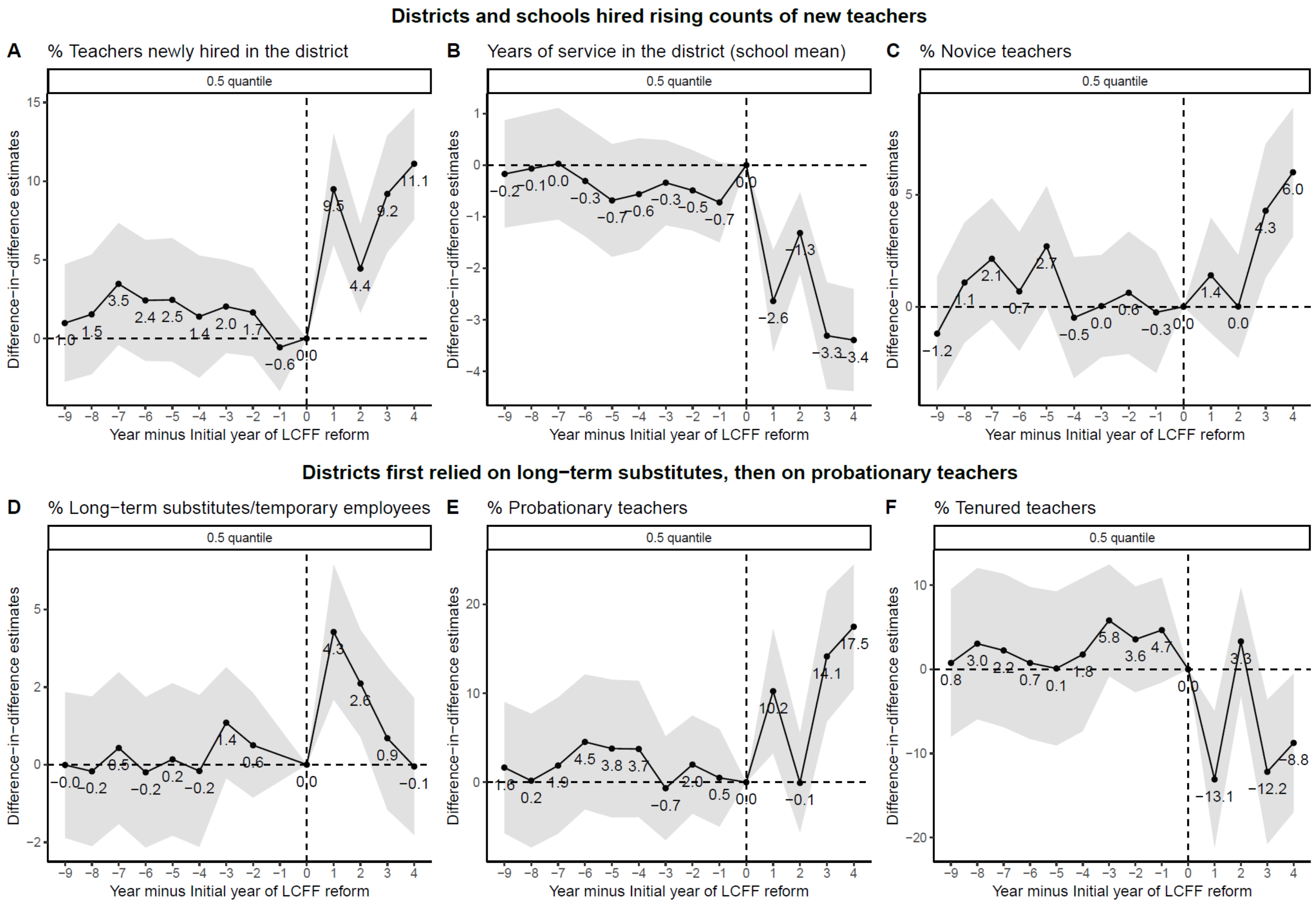

Figure 3.7: Event study estimates of Local Control Funding effects on school-level distributions of teacher characteristics (median, 0.5 quantile).

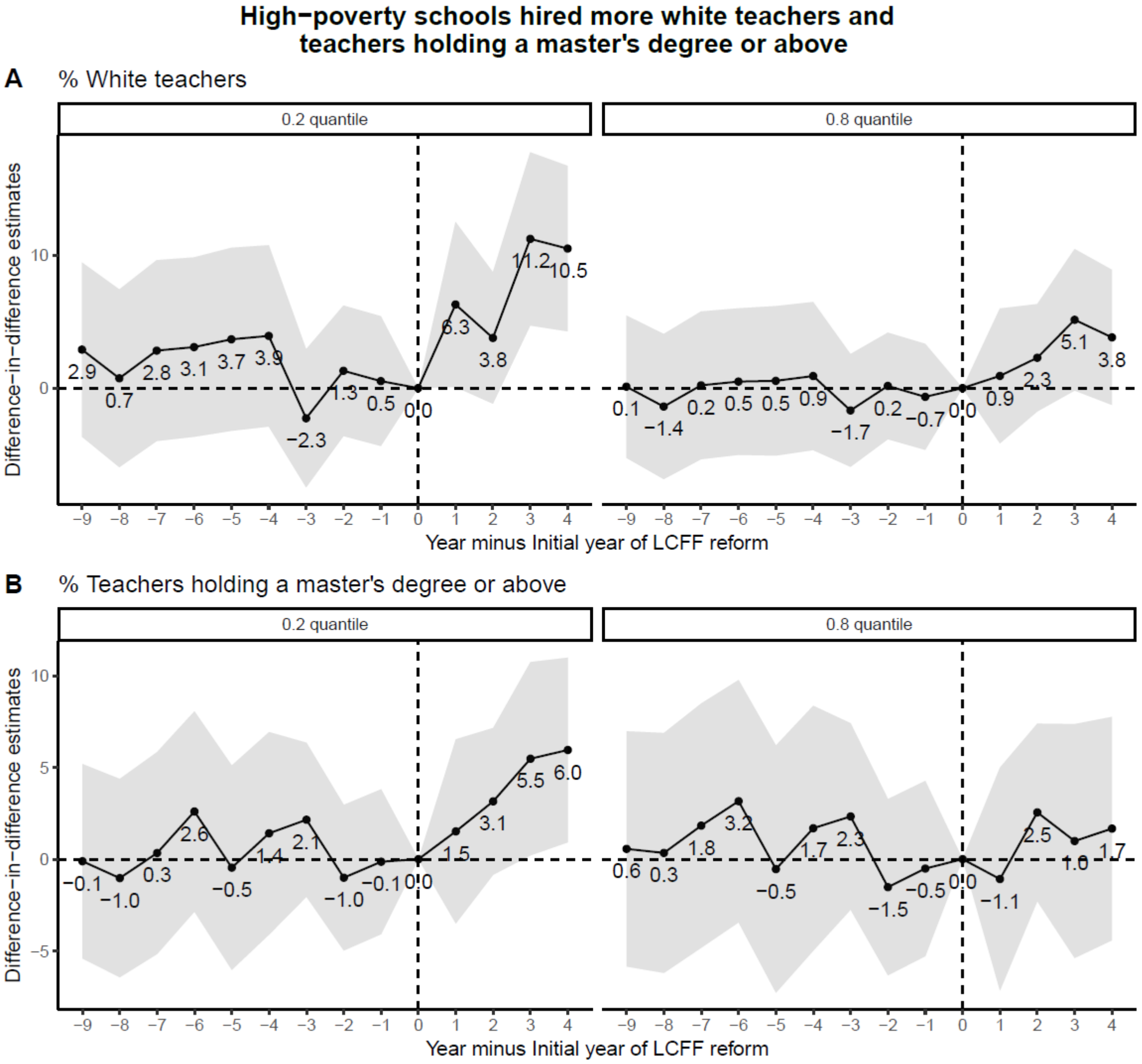

Figure 3.8: Event study estimates of Local Control Funding effects on school-level distributions of shares of white teachers and master’s degree holders (0.2 and 0.8 quantiles).

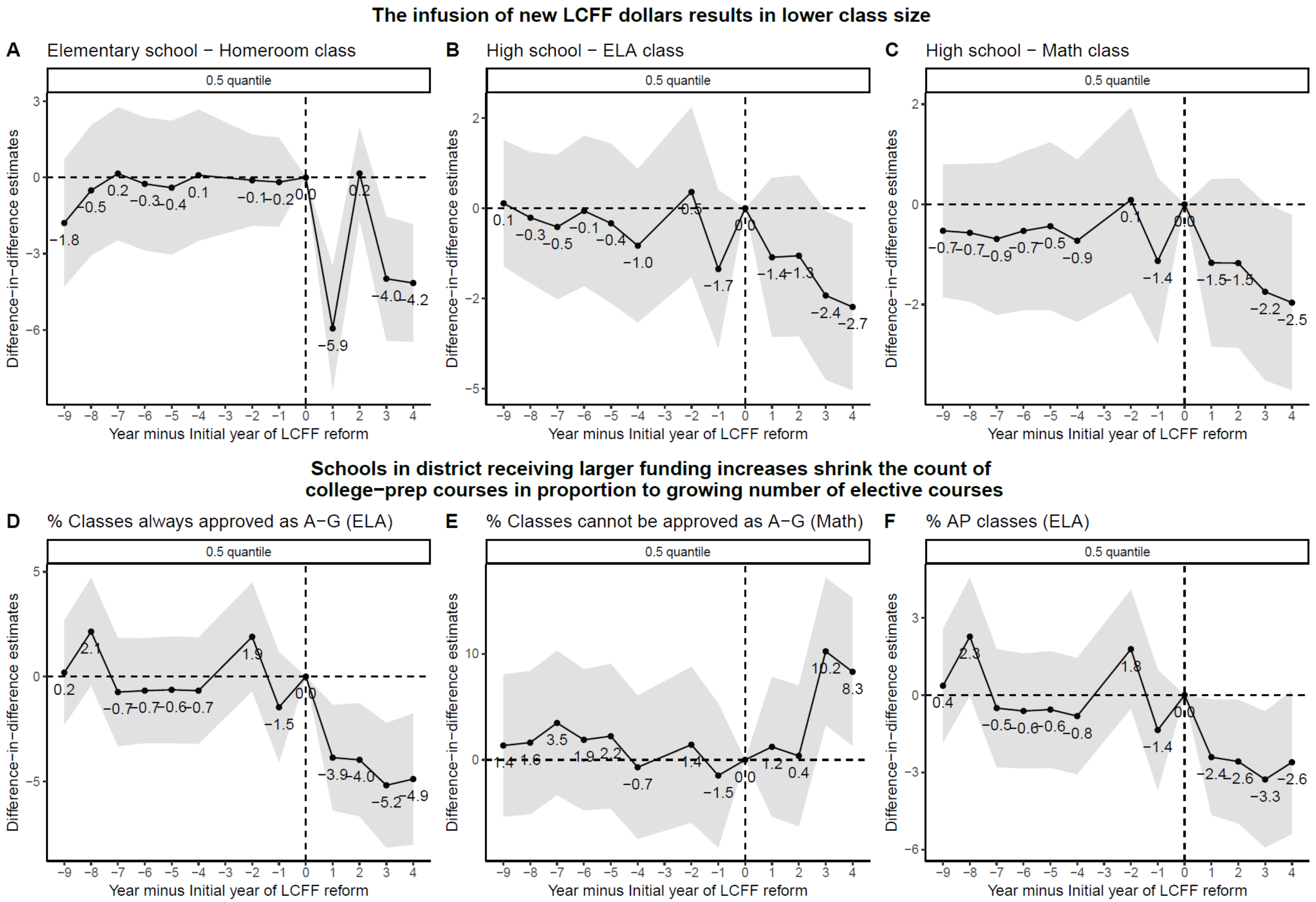

Figure 3.9: Event study estimates of Local Control Funding effects on distributions of school-level organizational practices and curricular.

We also found that LCF-induced funding increases lowered average class size at elementary and high school levels (panels A, B, and C in Figure 3.9). Elementary schools show immediate declines in class size after one year of LCF exposure, while high schools felt the effect incrementally as the exposure to the reform accumulates. Though statistically significant, effect sizes were modest: an LCF-induced 10% increase in funding reduced the elementary homeroom class size by 0.59 in the first year and high school ELA and math class sizes by 0.27 and 0.25, respectively, after four years of exposure. We found little evidence that spending increases significantly altered the count of instructional periods assigned to teachers each day.

Panels D, E, and F display how the curricular structure changed among high schools in the wake of LCF-induced changes in spending. An LCF-induced 10% increase in per-pupil spending significantly lowered the percentage of all high school ELA classes that qualified for the college-prep A-G designation by about 0.52 percentage point in the third year of reform. Shares of ELA and math Advanced Placement classes declined following the exogenous spending increases.

DiD estimates for each quantile presented in Table 3.3 suggest that the proportional shrinkage of rigorous college-prep courses (AP and A-G) occurs most severely at the higher quantile ($\tau = 0.8$), and in schools that began with higher shares of college -prep courses in the reference year (low-poverty schools). These results present some evidence in favor of the hypothesis that schools in districts receiving larger funding increases ended up reducing presence of college-prep courses relative to a growing number of elective courses.

We found little evidence that LCF-induced funding successfully reduced disparities in access to experienced teachers or A-G courses by EL students. We found that ELs' access to experienced teachers worsened for math classes in high-poverty schools, which showed greater inequities in the reference year ($\tau = 0.8$). We infer that the high-poverty schools suffering from high teacher turnover rates filled vacancies with novice or inexperienced teachers, then assigned them to classes with lower-achieving students.

We do see that these inequities in the curricular structure and teacher assignments persist from pre-LCF periods and widen particularly in high-poverty schools during the post-reform period. This raises concern that *within-school* sorting may prevent experienced teachers from being assigned to students who need them most, even when *between-school* teacher sorting can be mitigated. The steady expansion of elective courses, following the arrival of LCF dollars, may prevent low-achieving students from pursuing a more rigorous curriculum.

3.7 Discussion and policy implications

California's effort to progressively fund districts serving poor children, while decentering fiscal control to local districts, offers a remarkable policy experiment. Our findings demonstrate how resulting spending gains far exceeded levels predicted, due to the LCF reform and the state's set-aside for K-12 funding during a robust economy. The state's distribution of funding among local districts changed dramatically, favoring those hosting larger shares of disadvantaged pupils (relative to district counter-factual levels) as intended by policy makers.

Still, this sizeable infusion of new funding unfolded under a hazy theory-of-action in terms of what inputs (resource patterns) and organizational practices might change, likely mediators that operate proximal to student learning. Five years into California's ambitious funding reform – boosting yearly K-12 spending by $23 billion – we observed little discernible progress in narrowing student achievement gaps statewide. Our analytic strategy then endeavored to understand how LCF altered the mix of teacher inputs and organizational practices in ways that failed to lift the lowest achieving pupils

Our findings do show that new funding went for additional teaching positions, modestly lowering class sizes, while yielding organizational effects that may have worked against greater equity within schools. We also find that socioeconomic conditions under which funding infusions arrive matter. That is, the descriptive trends in the post-2013 period differed by the wealth or poverty of families served by schools. Many districts, quickly searching for new teachers as funding arrived, initially relied heavily on long-term substitutes or temporary employees, and then on rising shares of novice and probationary (non-tenured) teachers. And schools in high-poverty conditions relied even more heavily on novice teachers, compared with their better-off counterparts.

Key features of school organizations shifted modestly as well. High-poverty schools experienced a modest class-size decline from 24.4 to 23.3 pupils; low-poverty schools saw mean class size fall from 25.6 to 24.0. The count of teaching periods assigned to staff ticked up slightly in high-poverty schools, 4.0 to 4.1, while falling in the economically best-off quintile of schools, from 4.0 to 3.9. These small magnitudes – while significantly affected by the exogenous portion of the LCF reform – reveal that basic organizational features of schools and teachers' working conditions did not change dramatically on average. Much of the new funding may have simply returned schools to their pre-recession staffing levels.

On the other hand, we observed consequential shifts in teaching assignments and curricular structuring in schools located in districts that received stronger finance gains, organizational changes that likely worked *against* the state's pro–equity intentions and efforts by schools to narrow achievement gaps. First, as many districts enjoyed fresh funding and hired novice teachers, these less experienced staff were assigned to classes with higher shares of EL pupils.

Between the year just prior to LCF (2012-13) and 2017-18, the excess share of ELs in classes taught by probationary teachers, compared with tenured teachers, climbed from 4.7% to 10.7% in high-poverty high schools, while declining in the best-off quintile of schools (2.4% to 1.7%).

Second, the curricular structure of high schools began to de-emphasize college-prep courses after enactment of LCF. This may have occurred in response to the partial collapse of No Child Left Behind and its stricter testing-and-accountability policies. Or, the hiring of new teachers may have allowed schools the chance to recover lost elective courses. For whatever reason, the simple index of A-G college-prep classes (percentage A-G minus non-A-G) remained constant for low-poverty high schools, while falling 3.4 percentage points in math and 10.2 points in ELA among the poorest quintile of school after LCF implementation.

Overall, these trends suggest that LCF's large infusion of new funding has shifted the average profile of staff toward less experienced teachers, lowered class size, and moved high school curricula away from a college-prep emphasis. Future work should examine whether these trends persist, or whether local education leaders can retain more experienced teachers and enrich the curricular structure and other internal workings of schools that operate proximally to student learning.

The exogenous portion of the LCF reform helped to explain other modest changes in teacher attributes and organizational features of schools. For the middle quintile of schools in the event study, we observed significant treatment effects: a short-term spike in the hiring of long-term substitutes, followed by heightened reliance on novice teachers. The surge in hiring exercised by schools with low shares of white teachers (higher-poverty schools) led to increased employment of white teachers and holders of master's degrees.

The event study also revealed LCF-treatment effects in shrinking average class size for elementary and high schools. But the infusion of new spending also led high schools to reduce the proportion of classes qualifying for A-G designation, along with declines in the share of Advanced Placement courses. Some educators may applaud the robust return of electives. But the shift away from college-prep offerings may fail to boost college-going or to narrow achievement disparities among high school graduates.

Our study holds certain limitations. The focus on teacher characteristics and organizational features of schools is limited by available data for the 15-year time series. These mediators make intuitive sense and illuminate how finance infusions do indeed alter historical trends in teacher inputs, key facets of school organizations, curricular patterns and teaching assignments. Still, we have much to learn about the predictive validity of these particular mediators in terms of shaping student learning over time. Ideally, other mediators that operate between finance infusions and achievement will become available over time.

We are presently building district-by-district data on teacher salary levels, which have variably shifted upward over the same period. It may be that large finance infusions contribute to wage and fringe-benefit gains, leaving comparatively thin resources that actually reach classrooms or school principals.

California's education department maintains quite rich course-level data, which we exploited to tease apart how novice teachers are assigned to courses dominated by EL students, the pattern worsened by rapid infusion of new teacher hiring. Still, much work remains in understanding the racial or class-based tracking of students within schools. The arrival of fresh inputs and new dollars could be deployed to reduce (or reinforce) internal tracking of low and higher-achievement students inside schools, again depending on organizational practices. Our results suggest that medium-term reliance on less experienced teachers, who are then assigned to classes with higher shares of EL pupils, fails to address inequities in the opportunity to learn. It's a ripe example of how gains in dollars or raw inputs are variably mediated by how principals and school-level actors shape the social organization of their schools.

Finally, former Gov. Jerry Brown displayed little interest in learning about the effectiveness of his massive finance reform. This may change as LCF comes under greater scrutiny by a recently elected governor and new legislative leaders. Despite calls in policy and scholarly circles for learning not simply about *whether* money matters, but also when or *through what mediators*, it remains a question that's under-theorized and rarely examined with longitudinal data in the wake of major finance reforms. This limits our understanding of when – through which organizational practices and for which schools – more money likely matters.

Knowledge continues to grow regarding how the organization of schooling affects social cohesion and motivates students and teachers (e.g., Bryk et al., 2010). But little has been learned about how such mediators – be they fresh mixes of teacher inputs or textured organizational practices – are touched by sizeable infusions of new funding. When the political stars do align to progressively fund schools, we must rigorously dig into whether these initiatives truly advance fairness and, if so, through what changes in staffing patterns and the social organization of schools.

Bibliography

Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, *70*(1), 91-117.

Angelis, D. D., Hall, P., & Young, G. A. (1993). Analytical and bootstrap approximations to estimator distributions in L_1 regression. *Journal of the American Statistical Association*, *88*(424), 1310-1316.

Antonelli, J., Trippa, L., & Haneuse, S. (2016). Mitigating bias in generalized linear mixed models: The case for Bayesian nonparametrics. *Statistical science, 31*(1), 80.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, 1152-1174.

Augenblick, J., Myers, J., Anderson, A. (1997). Equity and adequacy in school funding. *Future of Children, 7*, 63-78.

Bassett Jr, G., & Koenker, R. (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, *77*(378), 407-415.

Basu, S., & Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association*, *98*(461), 224-235.

Bersin, A., Kirst, M.W., & Liu, G. (2008). Getting Beyond the Facts: Reforming California School Finance. The Chief Justice Earl Warren Institute on Race, Ethnicity, and Diversity. University of California, Berkeley, CA. Retrieved from https://www.law.berkeley.edu/files/GBTFissuebriefFINAL.pdf.

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). *Experimental evidence on distributional effects of Head Start* (No. w20434). National Bureau of Economic Research.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817-842.

Bloom, H. S., & Weiland, C. (2015, March). *Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study*. New York, NY: MDRC.

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 518-540.

Brown, J. (2013). Governor Brown signs historic school funding legislation. Sacramento: Office of the Governor. Access at: https://www.gov.ca.gov/news.php?id=18123.

Bruno, P. (2018). District dollars 2: California school district finances, 2004-5 through 2016-17. Technical report. Getting Down to Facts II. Retrieved from: https://gettingdowntofacts.com/sites/default/files/2018-09/GDTFII_Report_Bruno.pdf.

Bryk, A., Sebring, P., Allensworth, E., Luppescu, S., & Easton, J. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago: University of Chicago Press.

Cade, B. S., Noon, B. R., & Flather, C. H. (2005). Quantile regression reveals hidden bias and uncertainty in habitat models. *Ecology*, *86*(3), 786-800.

CDE, California Department of Education (2018a). Annual Financial Data. Retrieved from https://www.cde.ca.gov/ds/fd/fd/

CDE, California Department of Education (2018b). The California Longitudinal Pupil Achievement Data System, staff date file. Retrieved from https://www.cde.ca.gov/ds/sd/df/

California Office of the Legislative Analyst (2013). *An overview of the Local Control Funding Formula*. CA: Sacramento.

California Office of the Legislative Analyst (2018). *2019-20 Budget: Proposition 98 Outlook*. CA: Sacramento.

Candelaria, C. A., & Shores, K. A. (2019). Court-ordered finance reforms in the adequacy era: Heterogeneous causal effects and sensitivity. *Education Finance and Policy, 14*(1), 31-60.

Card, D., & Krueger, A. B. (1992). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of political Economy, 100*(1), 1-40.

Card, D., & Payne, A. A. (2002). School finance reform, the distribution of school spending, and the distribution of student test scores. *Journal of Public Economics, 83*(1), 49-82.

Chambers, J., Shambaugh, L., Levin, J., Muraki, M., & Poland, L. (2008). *A tale of two districts: A comparative study of student-based funding and school-based decision making in San Francisco and Oakland Unified School Districts*. Washington, DC: American Institutes for Research.

Chetverikov, D., Larsen, B., & Palmer, C. (2016). IV quantile regression for group-level treatments, with an application to the distributional effects of trade. *Econometrica, 84*(2), 809-833.

Chivers, C. (2012). *MHadaptive: General Markov chain Monte Carlo for Bayesian inference using adaptive Metropolis-Hastings sampling*. R package version 1.1-8. http://CRAN.R-project.org/package=MHadaptive.

Choi, H. M., & Hobert, J. P. (2013). Analysis of MCMC algorithms for Bayesian linear regression with Laplace errors. *Journal of Multivariate Analysis*, *117*, 32-40.

Conlon, E. M., & Louis, T. A. (1999). Addressing multiple goals in evaluating region-specific risk using Bayesian methods. In A. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J. F. Viel, & R. Bertollini (Ed), *Disease Mapping and Risk Assessment for Public Health* (pp. 31–47). Chichester: Wiley.

Congdon, P. D. (2019). *Bayesian Hierarchical Models: With Applications Using R (2nd edition)*. CRC Press.

Cornman, S., Zhou, L., Howell, M., & Young, J. (2018). *Revenues and expenditures for public elementary and secondary education: School year, 2014-15*. Washington, DC: National Center for Education Statistics.

Davino, C., Furno, M., & Vistocco, D. (2014). *Quantile regression: Theory and applications*. New York: Wiley.

Department of Education (2018). State schools chief Tom Torlakson announces results of California assessment of student performance and progress on tests. Sacramento, October 2.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials, 7*(3), 177-188.

Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L., & Jordan, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics, 64*(2), 635-644.

Dorazio, R. M. (2009). On selecting a prior for the precision parameter of Dirichlet process mixture models. *Journal of Statistical Planning and Inference*, *139*(9), 3384-3390.

Dunson, D. B., Pillai, N., & Park, J. H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 163-183.

Edmunds, J. A., Unlu, F., Glennie, E., Bernstein, L., Fesler, L., Furey, J., & Arshavsky, N. (2017). Smoothing the transition to postsecondary education: The impact of the early college model. *Journal of Research on Educational Effectiveness*, *10*(2), 297-325.

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika*, *103*(1), 1-20.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577-588.

Fensterwald, J. (2015). Torlakson reinterprets department's stance on teacher raises. *EdSource*, June 15. Access at: https://edsource.org/2015/torlakson-reinterprets-departments-stance-on-teacher-raises/81528.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28*(2), 465-498.

Gelfand, A. E., & Kottas, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *11*(2), 289-305.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis (3rd Edition)*. CRC press.

Ghidey, W., Lesaffre, E., & Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, *60*(4), 945-953.

Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, *87*(418), 533-540.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *159*(3), 385-409.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of educational research, 66*(3), 361-396.

Guryan, J. (2001). *Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts* (NBER Working Paper No. 8269). Cambridge, MA. https://doi.org/10.1007/s13398-014-0173-7.2

Hanushek, E. (2016). What matters for student achievement. *Education Next, 16*(2), 18-26.

Hanushek, E., & Woessmann, L. (2017). School resources and student achievement: A review of cross-country economic research. pp. 149-170 in *Cognitive abilities and education outcomes*, edited by M. Rosen. New York: Springer.

Hao, L., & Naiman, D. Q. (2007). *Quantile regression* (No. 149). Thousand Oaks: Sage.

Harwell, M., Kohli, N., & Peralta, Y. (2017). Experimental design and data analysis in computer simulation studies in the behavioral sciences. *Journal of Modern Applied Statistical Methods, 16*(2), 2.

Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica*, 1377-1398.

Hausman, J., Luo, Y., & Palmer, C. (2014). Errors in the dependent variable of quantile regression models. *Working paper*. Department of Economics, MIT.

Hendricks, W., & Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American statistical Association*, *87*(417), 58-68.

Hyman, J. (2017). Does money matter in the long run? Effects of school spending on educational attainment. *American Economic Journal: Economic Policy, 9*(4), 256-80.

Ishwaran, H., & Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, *87*(2), 371-390.

Jackson, C. K., Johnson, R. C., & Persico, C. (2015). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *The Quarterly Journal of Economics, 131*(1), 157-218.

Johnson, R., & Tanner, S. E. A. N. (2018). *Money and freedom: The impact of California's school finance reform on academic achievement and the composition of district spending*. Technical report. Getting Down to Facts II. Retrieved from: https://gettingdowntofacts.com/sites/default/files/2018-09/GDTFII_Report_Johnson.pdf.

Katz, L. F., Kling, J. R., & Liebman, J. B. (2000). *Moving to opportunity in Boston: Early results of a randomized mobility experiment* (No. w7973). Cambridge, MA: National Bureau of Economic Research.

Klopfer, J. (2017). Labor supply, learning time, and the efficiency of school spending: Evidence from school finance reforms. Working Paper. Retrieved from: https://drive.google.com/file/d/1ZkRth2iXI_uB9aMhfHH_GvdTkyxe_QFJ/view.

Kohli, S. (2016). How did LAUSD spend $450 million? Not on the high-needs students the money was for, state officials say. *Los Angeles Times*, August 9. Access at: http://www.latimes.com/local/education/la-me-edu-lausd-state-funding-decision-20160807-snap-story.html.

Koenker, R. & d'Orey (1994). Computing regression quantiles. *Applied Statistics, 43*, 410–414.

Koenker, R. (2005). *Quantile regression* (No. 38). Cambridge: Cambridge University Press.

Koenker, R. (2016). *quantreg: Quantile Regression*. R package version 5.29. http://CRAN.R-project.org/package=quantreg.

Kozumi, H., & Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of statistical computation and simulation*, *81*(11), 1565-1578.

Lafortune, J., Rothstein, J., & Schanzenbach, D. W. (2018). School finance reform and the distribution of student achievement. *American Economic Journal: Applied Economics, 10*(2), 1-26.

Lafortune, J. (2019). *School Resources and the Local Control Funding Formula: Is Increased Spending Reaching High-Need Students?*. Public Policy Institute of California, San Francisco.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*(364), 805-811.

Leslie, D. S., Kohn, R., & Nott, D. J. (2007). A general approach to heteroscedastic linear regression. *Statistics and Computing*, *17*(2), 131-146.

Levin, J., Chambers, J., Epstein, D., Mills, N., Archer, M., Wang, A., & Lane, K. (2013). *Evaluation of Hawaii's weighted student formula*. San Mateo, CA: American Institutes for Research.

Liu, J., & Dey, D. K. (2008). Skew random effects in multilevel binomial models: an alternative to nonparametric approach. *Statistical modelling*, *8*(3), 221-241.

Lockwood, J. R., Castellano, K. E., & Shear, B. R. (2018). Flexible Bayesian models for inferences from coarsened, group-level achievement data. *Journal of Educational and Behavioral Statistics*, *43*(6), 663-692.

Loeb, S., Grissom, J., & Strunk, K. (2006). *District dollars: Painting a picture of revenues and expenditures in California's school districts.* Stanford: Manuscript.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, *79*(386), 393-398.

MacEachern, S. N., & Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, *7*(2), 223-238.

Malen, B., Dayhoff, J., Egan, L., & Croninger, R. G. (2017). The challenges of advancing fiscal equity in a resource-strained context. *Educational Policy, 31*(5), 615-642.

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the shape of a random-effects distribution: why getting it wrong may not matter. *Statistical science*, 388-402.

Meager, R. (2016). *Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature*. Manuscript: MIT.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, *11*(1), 57-91.

Miles, K. H., & Roza, M. (2006). Understanding student-weighted allocation as a means to greater school resource equity. *Peabody Journal of Education, 81*(3), 39-62.

Miratrix, L., Feller, A., Pillai, N., & Pati, D. (2016). Using Dirichlet Processes for Modeling Heterogeneous Treatment Effects across Sites. *Abstracts of 2016 Society for Research on Educational Effectiveness Annual Conference.* https://files.eric.ed.gov/fulltext/ED567507.pdf.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133-161.

Murugiah, S., & Sweeting, T. (2012). Selecting the precision parameter prior in Dirichlet process mixture models. *Journal of Statistical Planning and Inference, 142*(7), 1947-1959.

Normand, S. L. T., & Shahian, D. M. (2007). Statistical and clinical aspects of hospital outcomes profiling. *Statistical Science*, *22*(2), 206-226.

Odden, A., & Picus, L. (2014). *School finance: A policy perspective*. New York: McGraw-Hill.

Otsu, T. (2008). Conditional empirical likelihood estimation and inference for quantile regression models. *Journal of Econometrics*, *142*(1), 508-538.

Paddock, S. M., Ridgeway, G., Lin, R., & Louis, T. A. (2006). Flexible distributions for triple-goal estimates in two-stage hierarchical models. *Computational statistics & data analysis*, *50*(11), 3243-3262.

Papke, L. E. (2008). The effects of changes in Michigan's school finance system. *Public Finance Review, 36*(4), 456-474.

Partnership for Los Angeles Schools (2017). *Making equity the foundation for L.A. Unified budgeting. Los Angeles.* Policy Brief. Retrieved from https://partnershipla.org/wp-content/uploads/2018/07/Making-Equity-the-Foundation-of-LA-Unified-Budgeting.pdf.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*(2), 287-312.

Pinheiro, J. C., Liu, C., & Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, *10*(2), 249-276.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics*, *32*(1), 143-155.

R Core Team. (2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. http://www.R-project.org/.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, *3*(3), 215-232.

Rabe-Hesketh, S., & Skrondal, A. (2012). Multilevel and Longitudinal Modeling Using Stata. *Stata Press books*.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*(2), 75-98.

Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307-335.

Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475-499.

Reed, C., & Yu, K. (2009). A Partially collapsed Gibbs sampler for Bayesian quantile regression. http://bura.brunel.ac.uk/handle/2438/3593.

Reich, B. J., Bondell, H. D., & Wang, H. J. (2010). Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics*, *11*(2), 337-352.

Reich, B. J., Fuentes, M., & Dunson, D. B. (2012). Bayesian spatial quantile regression. *Journal of the American Statistical Association*, *106*(493), 6–20.

Reich, B. J., & Smith, L. B. (2013). Bayesian quantile regression for censored data. *Biometrics*, *69*(3), 651-660.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, *6*(4), 377-401.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Simon, H. (1979). Rational decision making in business organizations. *American Economic Review, 69,* 493-513.

Shen, W., & Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(2), 455-471.

Shen, W., & Louis, T. A. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *Journal of Computational and Graphical Statistics*, *8*(4), 800-823.

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

Spybrook, J. (2014). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *The Journal of Experimental Education*, *82*(3), 334-357.

Sriram, K., Ramamoorthi, R. V., & Ghosh, P. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Analysis*, *8*(2), 479-504.

Stefanski, L. A., & Carroll, R. J. (1990). Deconvolving kernel density estimators. *Statistics*, *21*(2), 169-184.

Syring, N., & Martin, R. (2015). Scaling the Gibbs posterior credible regions. *arXiv preprint arXiv:1509.00922*.

Tsionas, E. G. (2003). Bayesian quantile inference. *Journal of Statistical Computation and Simulation*, *73*(9), 659-674.

United Way and CLASS Coalition (2017). Incremental progress toward equity: Tracking the distribution of $3.8 billion LCF dollars in the Los Angeles Unified School District since 2013. Los Angeles: United Way of Greater Los Angeles.

Venturini, S., Dominici, F., & Parmigiani, G. (2015). Generalized quantile treatment effect: a flexible Bayesian approach using quantile ratio smoothing. *Bayesian Analysis*, *10*(3), 523-552.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*(433), 217-221.

West, M., Muller, P., & Escobar, M. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In P. Freeman & A. Smith (Eds.), *Aspects of Uncertainty: A Tribute to D. V. Lindley* (pp 363–386). Wiley.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 817-838.

White, I. R. (2010). simsum: Analyses of simulation studies including Monte Carlo error. *Stata Journal*, *10*(3), 369.

Wolf, R., & Sands, J. (2016). A preliminary analysis of California's new Local Control Funding Formula. *Education Policy Analysis Archives, 24*. Access at: http://files.eric.ed.gov/fulltext/EJ1100156.pdf.

Wright, D. L., Stern, H. S., & Cressie, N. (2003). Loss functions for estimation of extrema with an application to disease mapping. *Canadian Journal of Statistics*, *31*(3), 251-266.

Yang, Y., & He, X. (2012). Bayesian empirical likelihood for quantile regression. *The Annals of*

Statistics, *40*(2), 1102-1131.

Yang, Y., Wang, H. J., & He, X. (2015). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *International Statistical Review*, *84*(3), 327-344.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., ... & Paunesku, D. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364-369.

Yu, K., & Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, *54*(4), 437-447.

Yu, K., & Zhang, J. (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics—Theory and Methods*, *34*(9-10), 1867-1879.

Yu, K., & Stander, J. (2007). Bayesian analysis of a Tobit quantile regression model. *Journal of Econometrics*, *137*(1), 260-276.