Estimating the Effect of Florida's Low-100 Reading Program: Summarizing Regression Discontinuity Models with Bayesian Model Averaging

Seyfullah Tingir
Cambium Assessment

Russell Almond
Florida State University

Seyma Intepe-Tingir
University of St. Thomas

Abstract

In 2013, the state of Florida mandated an additional hour of intensive reading instruction for the lowest-performing 100 elementary schools across Florida. This requirement was implemented during the 2013–2014 academic year based on the schools' ranking in 2012–2013. This study assesses the effectiveness of the extra-hour intervention by using a regression discontinuity design (RDD). Often RDD analyses fit multiple models and then select a single best model using stepwise regression leading to overestimation of the effect size and underestimation of the standard error. This study used the Bayesian model averaging approach, which incorporates uncertainty about the best model. The estimated treatment effect, averaged over the six models and weighted by the models' posterior probabilities, is 6.1 points (d = .25) with a 95% confidence interval of 5.8 to 6.4 points.

Keywords: reading, intervention, regression discontinuity design, Bayesian model averaging

Reading ability is the foundation for learning in other academic areas (e.g., science, mathematics, and social science) and school success goes along with reading achievement (Lerner & Johns, 2012). Consequently, the most common reasons for failing to achieve academic success involve reading difficulties. Deficiencies in reading have been identified in 80% of students with learning disabilities (Rafdal et al., 2011). Fortunately, numerous large-scale intervention studies have indicated that intensive and systematic instruction can increase students' reading performance by approximately four to six percent (Torgesen, 2002). One straightforward way to increase academic achievement is to increase learning time, particularly reading time (Torgesen, 2000).

The National Reading Panel (NRP, 2000) reported five essential instructional components that are paramount to the successful acquisition of reading skills in young students: phonemic awareness, phonics, fluency, vocabulary, and comprehension. Students at risk for reading

difficulties might need more intensive interventions that include systematic and explicit instruction based on these critical elements. Trained staff should give instruction appropriate for the reading ability of the student. The frequency and duration of instruction are critical to increasing students' reading performance (Vaughn et al., 2012); therefore, increasing the learning time and giving more developmentally appropriate reading instruction should improve reading performance.

Reading instruction in the early elementary years is critical to building reading proficiency (Denton, 2012). Inadequate reading proficiency in early years negatively affects academic performance in later years (Lerner & Johns, 2012; Snow 2002; Snow et al., 2007). Early reading interventions are critical because reading problems persist through all areas of learning, and reading remediation becomes more difficult as children mature (Morocco, 2001; Torgesen et al., 2007).

Wanzek and Vaughn (2008) conducted two studies that examined student response to various levels of reading intervention for low-ability first-grade readers. Additional intervention time was implemented every day of the week for 50 days in total for the treatment group, which demonstrated enhanced learning compared to the control group. For first- to third-grade students, Denton et al. (2010) conducted research on word reading, fluency, and comprehension. They provided an additional intervention of 1–2 hours every day for a total of 8–16 weeks and found positive gains for treatment groups.

For the upper elementary grades, Ritchey et al. (2012) conducted an intervention for 123 fourth-grade students who were at risk for reading failure. The intervention was over 12 to 15 weeks with three additional 40-minute periods each week focusing on fluency, comprehension, vocabulary, and integrated text with science subject areas. The intervention group outperformed the control group significantly in terms of knowledge of science vocabulary and comprehension strategies. Students who were at higher risk benefited more than the low-risk group. Vadasy and Sanders (2008) examined the effectiveness of the reading program as an additional intervention for 119 fourth- and fifth-graders who performed below expectations on fluency. The intervention was conducted for an extra 30 minutes for four days per week for 18 weeks. The students using the reading program performed better than the students with regular curriculum in vocabulary and comprehension. Lastly, Torgesen et al. (2001) examined third- to fifth-grade students with reading challenges. Students were provided intensive intervention during two additional 50-minute sessions every day for a total of eight weeks. The treatment group increased their reading outcomes and achieved grade-level expectations. After two years, these students maintained their improvement in reading achievement. Overall, the literature review from previous years showed that additional instructional time is an effective method for at-risk students to improve their reading abilities.

Background: A Florida-Mandated Legislated Change

Maheady et al. (2006) emphasized that "powerful academic interventions can prevent and remediate reading failure before it leads to even more devastating outcomes" (p. 66). In line with this viewpoint, the state of Florida took a direct approach to this issue. To prevent reading failure, the Florida House passed Bill 5101 in 2012, known as the Low-100 program. This bill mandates an additional hour of reading instruction for the students in the lowest-performing 100 elementary schools with the lowest scores in reading, though the legislation did not provide additional funds to implement the program (Corbett, 2015). The aim of this requirement was to improve students' reading performance. The 100 lowest-performing elementary schools according to the state rating system were required to add an extra hour of reading instruction to their school day. In the next

school year (2014), this law was broadened to include the 300 elementary schools that had the lowest reading performance. In the original law, the additional hour must be provided by teachers or reading specialists, but the law modified the program by requiring the K–5 mentoring program to be implemented by a teacher who is effective at teaching reading. These schools were selected based on the lowest reading scores and changed every year.

The implementation of the policy in terms of the timing, approach, and training of the staff was left to the schools. The implementation times varied across schools: at the beginning of the day (7%), during the day (38%), at the end of the day (39%), and at different times on different days (15%). The additional hour also could be implemented as varied size of groups, guided instruction, or vocabulary study. Lastly, the professional training varied across the schools where 81% of the schools provided training and 45% of the schools provided additional planning time for teaching (West & Vickers, 2014). The Low-100 program required that the extra hour of instruction must consist of research-based instruction, varied instruction based on the student's proficiency, combinations of phonemic awareness, phonics, fluency, vocabulary, comprehension, guided practice, and integration with other subject areas (Corbett, 2015). Reading coaches, teachers, paraprofessionals, and volunteers were responsible for the extra intervention. Providing this program cost approximately $300,000–$400,000 for each school and $800 for each student. This requirement was implemented during the 2013–2014 academic year based on the school's ranking in 2012–2013. The program continued the next year for the lowest-performing 300 elementary schools.

General details about this program (Corbett, 2015):

- The 300 lowest-performing schools in reading are required to implement the supplemental academic instruction (SAI) school wide. The schools that implemented additional research-based reading instruction were the same in 2014–2015 and 2015–2016.
- These schools were selected based on having the lowest sums of points earned for reading achievement (Levels 3 and above) and reading learning gains as determined for the purpose of calculating school grades in 2013–2014.
- The intervention must be conducted by teachers or reading specialists who are proficient in teaching reading or by a K–5 mentoring program supervised by a teacher who is effective in teaching reading.
- Districts are responsible for identifying teachers who are effective in reading to supervise the extra hour of instruction.
- A survey must be conducted to report the documentation of expenditures at the end of the school year.
- Students who obtain a 2013–2014 FCAT Achievement Level 5 in reading may opt out of the program; all the other students must participate in the instruction unless they have special circumstances.
- Schools can use SAI designated funds for adjusting transportation to accommodate the extra instruction.
- Charter schools that fall in the 300 must implement SAI as well.
- The most common implementation of the SAI was to add the additional hour at the end of the school day.

Using information from the mandated policy change, this paper uses a regression discontinuity design (RDD) to examine whether the extra hour policy is effective in raising school average reading scores. Potentially, there are multiple models that could be used to estimate the effect size. This paper uses a Bayesian model averaging (BMA) approach to summarize across the candidate RDD models (Madigan & Raftery, 1994).

Methods

RDD is a type of quasi-experimental design that analyzes the difference between units (schools) above and below a cutoff score on an assignment variable (Shadish et al., 2002). The treatment is applied only to units on one side of the cutoff score. In this case, the cut score is implicit: based on the scores on the reading accountability measure between the 100th and 101st (or 300th and 301st) school. The RDD estimates the causal effect of the intervention after first adjusting for the variable used to assign the treatment (in this case the accountability scores for reading). The assumption is that the effect of the assignment variable is a smooth function, but there are several possible smooth functions that could be used. In a standard RDD approach, this introduces researcher degrees of freedom that can be used to maximize the effect size based on the observed data.

The Florida Department of Education (FLDOE) implemented the extra-hour intervention based on a ranking of the schools in 2012–2013. The FLDOE ranked the schools by summing two scores, based on the annual state reading tests: "the percent of students scoring satisfactory or higher" and "reading points for gains score," both obtained as part of the school's annual accountability examination (FLDOE 2013, 2014). The percent of students scoring a satisfactory or higher score was calculated by taking the percent of students who scored satisfactory in reading by scoring at or above FCAT 2.0 Achievement Level 3 in reading or by scoring at or above Performance Level 4 in reading on the Florida Alternate Assessment (FAA). The reading points for gains score was calculated by taking the percent of students making learning gains (gaining 5 points or more on the FCAT or maintaining a Level 3 [proficient] or higher) in reading, with additional weighting for students moving to Level 4 or 5 from a lower achievement level on the FCAT 2.0 or End-of-Course (EOC) assessments and for previously low-performing students making greater-than-expected gains on state assessments, including the FAA (Stewart, 2014).

The 100 schools scoring the lowest on this metric (the cutoff was 90) were placed into the Low-100 intervention for the RDD research model (see Figure 1). The schools below this cutoff score (the lowest 100) were in the treatment group while the remaining schools (n = 1,671) formed the control group. Table 1 shows the descriptive statistics for the reading scores of the two years. Figure 1 depicts the distributions of the reading scores for two years.

Table 1. *Descriptive statistics for reading scores*

Group	Year	*M*	*SD*	Minimum	Maximum
Treatment (n = 89)	2012–13	79.8	8.55	50.0	89.0
	2013–14	94.4	15.07	54.0	133.0
Control (n = 1,671)	2012–13	125.9	19.63	90.0	183.0
	2013–14	128.6	21.82	63.0	185.0

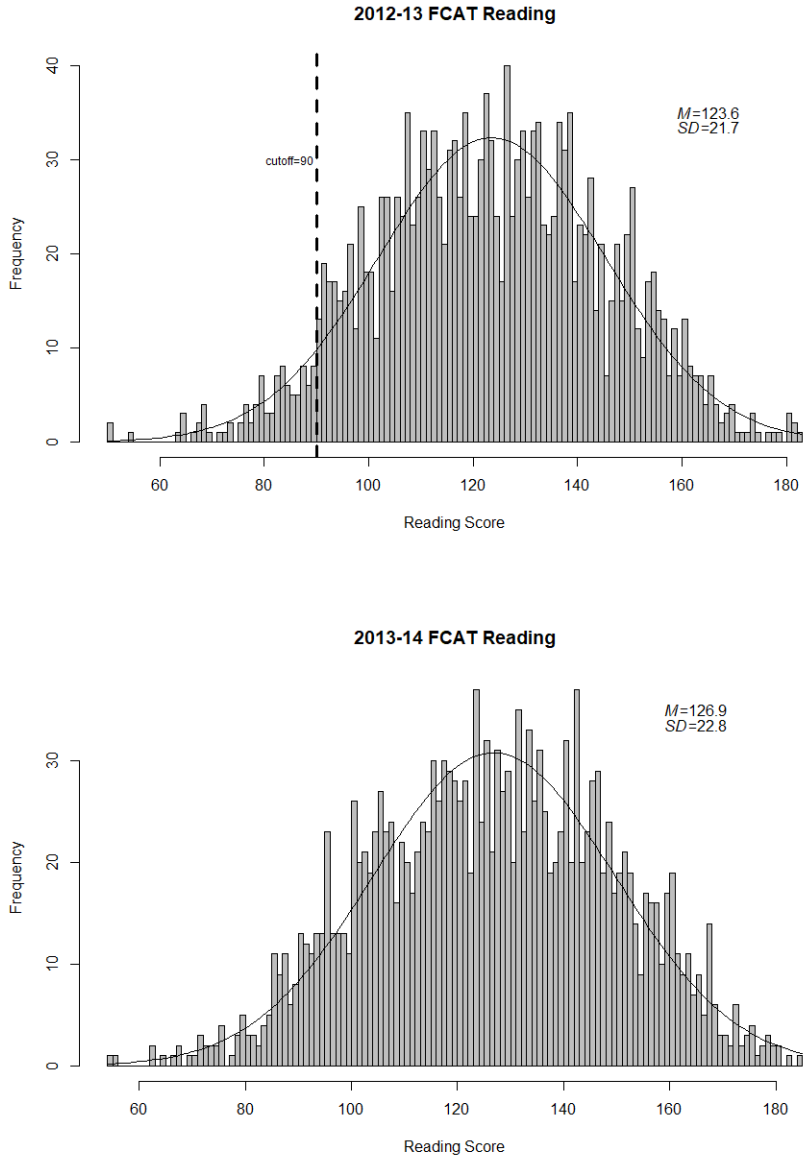

Figure 1. Distributions of the reading scores in 2012–13 and 2013–14

A simple linear RDD is just a linear model with a covariate for the assignment variable:

$$Y = \alpha + \beta_0 T + \beta_1 (X - c) + \varepsilon, \quad (1)$$

where Y is the outcome, T is the treatment indicator (T=1 for treatment, T=0 for control), X is the assignment variable, c is the cutoff score, and $(X - c)$ centers the distribution at the cutoff. The coefficient β_0 is the effect of the treatment, the target of inference. If the pretest scores are centered at the cutoff score, X_{cc}, then Equation 1 becomes:

$$Y = \alpha + \beta_0 T + \beta_1 (X_{cc}) + \varepsilon. \quad (2)$$

Average Treatment Effect (ATE) is the difference in mean potential outcome scores of treatment schools and control schools at the cutoff. ATE is formally:

$$\beta_{0(ATE)} = E[Y_i^t] - E[Y_i^c], \quad (3)$$

where the expectation is taken over the whole sample. In the simple linear model, this is just β_0, but Equation 3 generalizes to more complicated models.

We used six RDD models, described below, to estimate the treatment effect. The standard analysis procedure would be to pick the model that fit the observed data best and use that to estimate the ATE. The standard procedure ignores any uncertainty about which model is best. The BMA technique summarizes across the models, incorporating model uncertainty into the standard error (Madigan & Raftery, 1994).

Data

The data set included all elementary schools in the state of Florida for which there were accountability scores for 2012–13 and 2013–14. The intervention schools were selected based on their performance on the sum of their reading achievement and learning gain scores in 2012–13. The extra-hour program was initially intended to select the lowest 100 schools. Since some of the schools were closed in 2013–14 and information on several schools in the 2013–14 school list was not clearly reported, data from only 89 treatment schools were included for all data analyses. We examined information from a total of 1,760 schools over both years.

Analysis

We used non-linear equations (i.e., quadratic and cubic) with and without interaction terms to explore overfitting and its role in decreasing the likelihood of bias (Shadish et al., 2002). Therefore, six RDD models were evaluated based on the ATE of each as shown in Table 2.

Table 2. *Regression Discontinuity Design Models*

Model 1	Linear	$Y_i = \alpha + \beta_0 T_i + \beta_1 X_{cc} + \varepsilon_i$
Model 2	Linear interaction	$Y_i = \alpha + \beta_0 T_i + \beta_1 X_{cc} + \beta_2 X_{cc} T_i + \varepsilon_i$
Model 3	Quadratic	$Y_i = \alpha + \beta_0 T_i + \beta_1 X_{cc} + \beta_2 X_{cc}^2 + \varepsilon_i$
Model 4	Quadratic interaction	$Y_i = \alpha + \beta_0 T_i + \beta_1 X_{cc} + \beta_2 X_{cc}^2 + \beta_3 X_{cc} T_i + \varepsilon_i$
Model 5	Cubic	$Y_i = \alpha + \beta_0 T_i + \beta_1 X_{cc} + \beta_2 X_{cc}^2 + \beta_3 X_{cc}^3 + \beta_4 X_{cc} T_i + \varepsilon_i$
Model 6	Cubic interaction	$Y_i = \alpha + \beta_0 T_i + \beta_1 X_{cc} + \beta_2 X_{cc}^2 + \beta_3 X_{cc}^3 + \beta_4 X_{cc} T_i + \beta_5 X_{cc}^2 T_i + \beta_6 X_{cc}^3 T_i + \varepsilon_i$

Figure 2 depicts the treatment and control group regression lines with cutoff score for the base model coefficients (Model 1). Dark gray squares represent treatment schools and gray triangles represent control schools. A centered cutoff score of zero is set for the discontinuity. The ATE of 8.003 can be visually observed at the cutoff score.

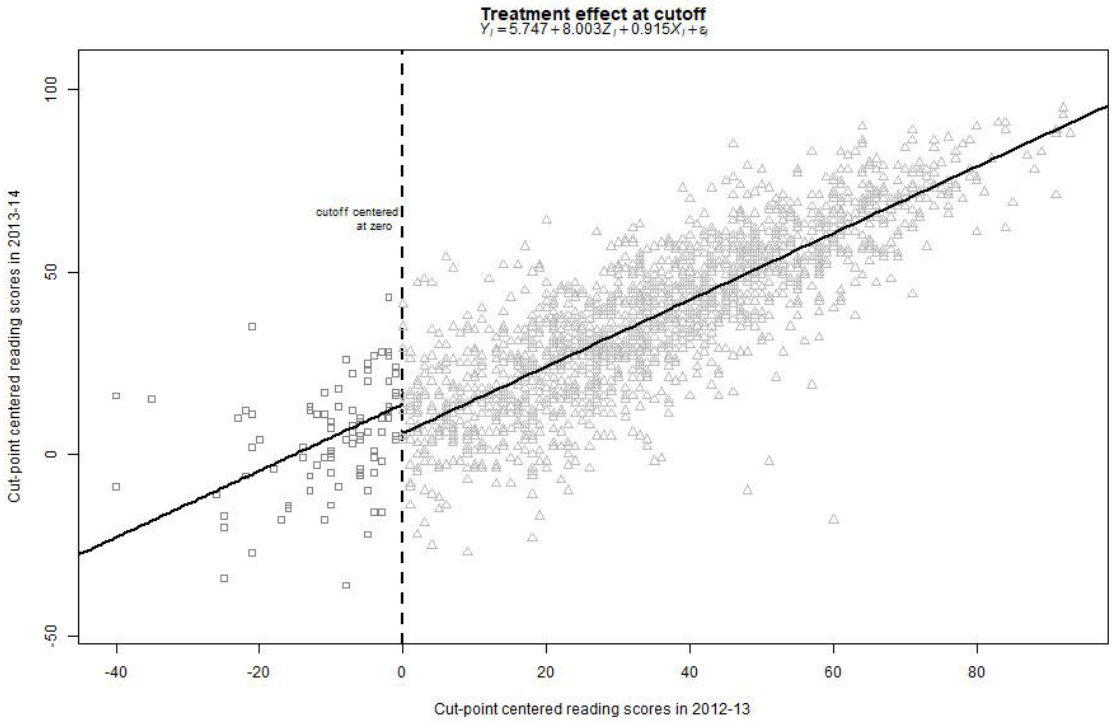

Figure 2. Regressions for treatment and control group

Table 3 shows the ATE estimates and standard errors. The ATE estimates from Model 4 and Model 5 are not significantly different from zero but the others are significantly different and positive. The standard stepwise regression procedure would find that the higher order terms in Models 2–6 were not significant and indicate that Model 1 was the best. The RDD analysis would then pretend that Model 1 was true and estimate the size of the treatment effect as *d* = *.35*. However, if one of Models 2–5 was closer to the truth, this would be an overestimate.

Table 3. *Average Treatment Effect (ATE) and Adjusted R-squared statistics*

	ATE	**Standard error**	**Lower limit (95%)**	**Upper limit (95%)**	**Effect size (Cohen's *d*)**	**Adjusted R-squared**
Model 1	8.003	1.526	5.008	10.998	0.35	0.702
Model 2	4.258	2.147	0.046	8.470	0.17	0.703
Model 3	5.723	1.953	1.891	9.555	0.25	0.703
Model 4	3.706	2.236	-0.681	8.093	0.16	0.703
Model 5	3.288	2.228	-1.082	7.659	0.14	0.706
Model 6	8.735	4.208	0.481	16.989	0.38	0.706

Model Averaging

Draper (1995) recommends dealing with this model uncertainty by averaging over the models. Madigan and Raftery (1994) use a weighted average using the posterior model probability as weights. This model averaging approach has been found to have better out-of-sample predictive accuracy than simply picking the best model and using that for predictions (Draper, 1995). The posterior model probability is proportional to the likelihood of the data under the model times the model prior probability. So, to use this method, we first must define prior model probabilities.

We chose the following prior probability for model M_k:

$$\Pr(M_k) = \gamma * \lambda^p, \tag{4}$$

where γ is the normalization constant and λ is a hyper-parameter that penalizes for extra parameters in the model, and p is the number of parameters in the model. Because the posterior distribution needs to be renormalized, the prior normalization constant can be set to one. There is less guidance on how to set λ. We choose to use values of .5, .75, and 1, which give a strong, medium, and no penalty at all for extra parameters in the model.

The likelihood of model k, $Pr(D|M_k)$, is just the probability of the observed data calculated using model k, with the parameters set to their maximum likelihood values. Lastly, the posterior distribution with a given D data is:

$$\Pr(M_k \mid D) \propto \Pr(D \mid M_k)\Pr(M_k). \tag{5}$$

Equation 5 must be normalized by dividing by the sum over all models considered, yielding posterior probabilities. These normalization values assign a weight (w) for each model:

$$w_k = \frac{\Pr(D \mid M_k)\Pr(M_k)}{\sum_{i=1}^{K} \Pr(D \mid M_i)\Pr(M_i)}. \tag{6}$$

Then, the summation of multiplications of each model's ATE (μ_k) and the model weight provides the weighted ATE ($\bar{\bar{\mu}}$):

$$\bar{\bar{\mu}} = \sum_{i=1}^{K} w_i \mu_i. \tag{7}$$

The weighted variance ($\sigma^2_{\bar{\bar{x}}}$) of ATE is a simple function of the individual model variances and the model ATEs:

$$\sigma^2_{\bar{\bar{x}}} = \sum_{i=1}^{K} w_i \sigma^2_{\bar{x},i} + w_i(\mu_i - \bar{\bar{\mu}})^2. \tag{8}$$

Note that Equation 8 has two terms. The first is the average of the squared standard error; this is the average within model variability. The second is the variance of the averages; this is a measure of between model variability. A big advantage of model averaging is that it incorporates this additional source of uncertainty when calculating standard errors, resulting in larger interval estimates.

Using the standard errors calculated using Equation 8, interval estimates can be formed by going two standard errors on either side of the average treatment effect estimate (Equation 7). Assuming that the posterior distributions for the treatment effect are roughly normal, this is an approximate

95% Bayesian credibility. This interval should also cover the true value approximately 95% of the time (classical confidence interval). In fact, intervals constructed using BMA are often better calibrated than classically constructed intervals, which ignore the uncertainty from model choice (Draper, 1995).

Results

Model Averaging

Table 4 shows the posterior distributions for the six models under the three different choices of prior. The $\lambda=1$ prior is essentially a non-informative prior and the last column shows the relative likelihoods of the six models. Although Model 1 has the highest likelihood, the likelihood is nearly identical for all six models. In Figure 2, there is a strong linear relationship between the scores in the two years in the control schools. Among the Low-100 schools, the relationship is close to linear, although there are some possible low outliers. Thus, the simple linear model (Model 1) fits fairly well. In fact, estimates of the additional parameters (i.e., quadratic and cubic and their interactions) in Models 2–6 are close to zero. Because the extra coefficients are close to zero in Models 2–6, the likelihood of the observed data under Models 2–6 is very close to that of Model 1. Even so, there are noticeable differences in estimated effect size (see Table 3). Given that all six models fit the data equally well, there is a natural preference for the simpler models. This is often expressed with a penalty term for model complexity common in model selection statistics like AIC and BIC (see Gelman et al., 2013, for definitions). In our example, the model complexity penalty is expressed through the prior distribution. Smaller values of λ represent stronger penalties. In Table 4, the first two columns show a preference for the simpler Model 1, but this preference is mostly based on the prior.

Table 4. *Model Averaging Results*

	Log-likelihood	**Posterior distribution (λ = .50)**	**Posterior distribution (λ = .75)**	**Posterior distribution (λ = 1)**
Model 1	-6930.126	0.410	0.263	0.1668
Model 2	-6927.058	0.205	0.197	0.1667
Model 3	-6928.379	0.205	0.197	0.1667
Model 4	-6926.667	0.102	0.148	0.1667
Model 5	-6917.914	0.051	0.110	0.1665
Model 6	-6916.209	0.025	0.083	0.1664

Table 5 shows the model averaging results as a function of λ. Smaller λ values put more weight on the simplest model (see Table 4). As Model 1 has one of the highest estimates for the average treatment effect, smaller values of λ lead to larger effect size estimates. The lower values of λ also have smaller variance, because the weight is more concentrated on a single model. Our preferred estimate is for the stronger prior on parsimony, thus our preferred estimate for the average effect of providing an extra hour of intervention to students at the Low-100 schools is an increase of 6.1 (d = .27*)* in the reading outcome scores, with a 95% confidence interval of 5.8 to 6.4 points.

Table 5. *Model Weighted Treatment Effects*

	Model weighted treatment effects	**Weighted variance**	**Lower limit (95%)**	**Upper limit (95%)**	**Effect size (Cohen's *d*)**
λ =0.50	6.10	7.29	5.76	6.44	0.27
λ =0.75	5.72	8.85	5.30	6.13	0.25
λ =1.00	5.62	10.79	5.11	6.12	0.25

Discussion

Programs like the Low-100, which revolve around targeted application of the treatment, are difficult to evaluate with conventional randomized designs. The regression discontinuity design offers a constructive approach to evaluating programs with systematic treatment assignment but does not offer a solution to the common problem of which model to select. Figlio et al. (2018) used a hierarchical single linear model within the RDD framework to control the variation at the school level but their analysis was still based on a single regression. Thus, estimates from a single best-fitting model will understate the uncertainty about the average treatment effect.

Note that simply choosing the single best model would produce an overestimate of the treatment effect (d = .35 instead of d = .27). Gelman and Loken (2013) call this phenomenon the "garden of forking paths," and note that performing model selection and then estimating the treatment effect with the same data will overstate the confidence of estimates, and often leads to positive bias in the effect size. Model averaging appears to reduce this problem; this is consistent with the findings of Draper (1995) where out of sample predictions from averages across several models performed better than predictions from the single "best" model.

Using RDD and model averaging, we estimated the average treatment effect of the Low-100 program at 6.1 (d = .27*)*, with a confidence interval of 5.8 to 6.4. According to Wiliam (2019), effect sizes should be interpreted according to the average yearly growth for the grade level. The FCAT assessment is given in Grades 3 through 5, which have an average growth between .6 and .4 standard deviations. So d = *.27* represents roughly one-half year's growth. However, the school scores are based on the percentages of students who are proficient or are making progress toward proficiency. This means that on average schools that applied the extra-hour intervention saw about 6% more of their students making progress toward or meeting their proficiency goal than they would have if they had not used the extra hour.

Note that 11 schools dropped out of the study, so this effect is effectively a treatment-on-treated estimate and not an intent-to-treat estimate (whose effect should be somewhat smaller). A second problem is that there is no uniformity in how the schools implemented the extra-hour program (West & Vickers, 2014) so this estimate provides little effective advice in how to implement such a program. Folsom et al. (2017) found in the lowest 300 schools analysis for the 2014–15 school year that students received indirect benefits, including perceived student gains. Their empirical analysis showed that the student gains occurred. Interviewees reported that those gains resulted from improvements in professional development and curricular and pedagogic changes, not necessarily from the extra hour of instruction.

Limitations and Future Work

Information from only 89 of the lowest 100 schools was available from the Florida Department of Education website, so the remaining 11 schools were excluded from our analysis. These schools are unlikely to be missing at random. In particular, the lowest-performing schools are much more likely to be closed or restructured than higher-performing schools. Thus, our estimate is of the treatment-on-treated effect rather than of the intent-to-treat analysis. We expect the latter effect would be somewhat smaller.

The only indicator for treatment we used was a single flag for membership in the Low-100 program; we did not have details on group size, the intensity of intervention, or the reading component focus for each grade at each school. West and Vickers (2014) did a survey and found considerable variability in the details of implementation. In some cases, this may have been good because it allowed the school to tailor the program to its unique needs. In other cases, the schools and districts might have benefited from guidance about what kinds of intervention would be the most effective. In regard to the extra-hour implementation time, for example, the schools that implemented it at the end of the day were the most successful (West & Vickers, 2014). All six of the models we considered used only two predictors, membership in the Low-100 and the previous year's FCAT scores, in different combinations. Additional covariates (in particular, school characteristics such as average size of classes, rates of poverty, and numbers of English language learners) added to the model might reduce the residual variance and lead to tighter interval estimates.

Despite (or maybe because of) the diversity of implementation, the extra hour of reading intervention in the Low-100 program was at least modestly successful. We would call for the program to be continued, although we would like to see more support at the state level both in terms of money for implementation and guidance as to the most effective implementation strategies.

The state of Florida expanded the Low-100 program to the lowest-performing 300 schools. This new program should be evaluated as well; however, there is a complication in that Florida changed is accountability test during the same year the Low-100 program was expanded to the Low-300 program. Thus, a major limitation is that we are unable to replicate this study with the next year's results. However, this is a temporary problem, and we think that RDD estimates using model averaging to summarize across models should continue to play a role in evaluating this program.

References

Corresponding Author: Seyfullah Tingir
Author Contact Information: seyfullah.tingir@gmail.com

Corbett, J. (2015). *Florida State Policy Brief: March 2015*. WestEd. https://files.eric.ed.gov/fulltext/ED559735.pdf

Denton, C. A. (2012). Response to intervention for reading difficulties in the primary grades: Some answers and lingering questions. *Journal of Learning Disabilities*, *45*(3), 232–243. https://doi.org/10.1177/0022219412442155

Denton, C. A., Kethley, C., Nimon, K., Kurz, T. B., Mathes, P. G., Shih, M., & Swanson, E. A. (2010). Effectiveness of a supplemental early reading intervention scaled up in multiple schools. *Exceptional Children*, *76*(4), 394–416. https://doi.org/10.1177/001440291007600402

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B, 57*, 45–70. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x

Figlio, D., Holden, K. L., & Ozek, U. (2018). Do students benefit from longer school days? Regression discontinuity evidence from Florida's additional hour of literacy instruction. *Economics of Education Review*, *67*, 171–183. https://doi.org/10.1016/j.econedurev.2018.06.003

Florida Department of Education. (2013, 2014) Florida School Accountability Reports. http://schoolgrades.fldoe.org

Folsom, J. S., Osborne-Lampkin, L., Cooley, S., & Smith, K. (2017). *Implementing the extended school day policy in Florida's 300 lowest performing elementary schools* (REL 2017-253). Regional Educational Laboratory Southeast. http://ies.ed.gov/ncee/edlabs

Gelman, A. & Loken, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking"' and the research hypothesis was posited ahead of time. (2013). [Unpublished paper] http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis.* Taylor & Francis.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*(428), 1535–1546. https://doi.org/10.1080/01621459.1994.10476894

Maheady, L., Mallette, B., & Harper, G. F. (2006). Four classwide peer tutoring models: Similarities, differences, and implications for research and practice. *Reading and Writing Quarterly*, *22*, 65–89. https://doi.org/10.1080/10573560500203541

Morocco, C. C. (2001). Teaching for understanding with students with disabilities: New directions for research on access to the general education curriculum. *Learning Disability Quarterly*, *24*(1), 5–13. https://doi.org/10.2307/1511292

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Pub. No. 00-4769). https://www.nichd.nih.gov/publications/pubs/nrp/smallbook

Rafdal, B. H., McMaster, K. L., McConnell, S. R., Fuchs, D., & Fuchs, L. S. (2011). The effectiveness of kindergarten peer-assisted learning strategies for students with disabilities. *Exceptional Children, 77*(3), 299–316. https://doi.org/10.1177/001440291107700303

Ritchey, K. D., Silverman, R. D., Montanaro, E. A., Speece, D. L., & Schatschneider, C. (2012). Effects of a tier 2 supplemental reading intervention for at-risk fourth-grade students. *Exceptional Children, 78*(3), 318–334. https://doi.org/10.1177/001440291207800304

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

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. RAND Corporation.

Snow, C. E., Porche, M. V., Tabors, P. O., & Harris, S. R. (2007). *Is literacy enough?* Brookes Publishing Co.

Stewart, P., 2014. *Guide to Calculating School Grades*. Retrieved December 14, 2015, from http://www.fldoe.org/accountability/accountability-reporting/publications-guides/

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice, 15*(1), 55–64. https://doi.org/10.1207/SLDRP1501_6

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. [Online Version]. *Journal of Learning Disabilities, 34*, 33–58. https://doi.org/10.1177/002221940103400104

Torgesen, J. K. (2002). Lessons learned from intervention research in reading: A way to go before we rest. In R. Stainthorp & P. Tomlinson (Eds.). *Learning and Teaching Reading: BJEP Monograph Series II: Psychological Aspects of Education–Current Trends* (pp. 89–104). British Psychological Society.

Torgesen, J., Houston D., Rissman, L., & Kosanovich, K. (2007). *Teaching all students to read in elementary school: A guide for principals*. RMC Research Corporation, Center on Instruction.

Vadasy, P. F., & Sanders, E. A. (2008). Benefits of repeated reading intervention for low-achieving fourth- and fifth-grade students. *Remedial and Special Education, 29*(4), 235–249. https://doi.org/10.1177/0741932507312013

Vaughn, S., Wanzek, J., Murray, C. S., Roberts, G. (2012). *Intensive interventions for students struggling in reading and mathematics: A practice guide*. RMC Research Corporation, Center on Instruction.

Vaughn, S., Wanzek, J., Murray, C. S., Scammacca, N., Linan-Thompson, S., & Woodruff, A. L. (2009). Response to early reading intervention examining higher and lower responders. *Exceptional Children, 75*(2), 165–183. https://doi.org/10.1177/001440290907500203

Wanzek, J., & Vaughn, S. (2008). Response to varying amounts of time in reading intervention for students with low response to intervention. *Journal of Learning Disabilities, 41*(2), 126–142. https://doi.org/10.1177/0022219407313426

West, M., & Vickers, B. (2014). *OPPAGA review of the Extra Hour Initiative*. The Florida Legislature's Office of Program Policy Analysis & Government Accountability. http://www.edweek.org/media/18florida-extra-hour-presentation.pdf.

Wiliam, D. (2019, April 4–8). *Why formative assessment is both generic and domain-specific*. [Paper presentation]. National Council on Measurement in Education Annual Meeting, Toronto, ON, Canada.