NEPC

NATIONAL EDUCATION POLICY CENTER

Review of *The Effects of Test-Based Retention on Student Outcomes Over Time: Regression Discontininuity Evidence from Florida*

Reviewed By

Joseph P. Robinson-Cimpian

University of Illinois at Urbana-Champaign

December 2015

Summary of Review

A recent NBER working paper examines Florida's policy to retain many low-scoring third graders. The report concludes that third-grade retention has immediate positive effects on the following year's test results, but these effects fade over the next six years, with no effect on graduation. The regression discontinuity methods used to estimate the effects, comparing students immediately above and below the law's cut-score, are generally good for making causal claims. But there is a serious shortcoming in the design—namely, the law requires that students below the cut-score receive intensive extra services intended to raise their subsequent achievement, and this applies to those retained and those promoted. This means the researchers do not know if these positive outcomes for those below the cut-score were due to the greater likelihood of retention or to the assurance of additional services. Also, two-thirds of students who fall below the cut-off score are nonetheless promoted because they fall into exception categories. Finally, the researchers exacerbate outcome differences by using an Instrumental Variable approach, which attributes the entire above/below-threshold difference to just retained students, effectively making the outcome difference appear more than three times as large. Because the policy stipulates that promoted students below the threshold also receive extra services that promoted students above the threshold do not receive, the IV approach is inappropriate. Even setting aside these problems, the study has extremely limited generalizability, restricted to students at or very near the threshold and directly affected by the policy. These and other problems call into serious question any causal claims of the longer-term effects and any policy utility for the report.

National Education Policy Center

School of Education, University of Colorado
Boulder, CO 80309-0249
Telephone: (802) 383-0058

Email: NEPC@colorado.edu
http://nepc.colorado.edu

This is one of a series of Think Twice think tank reviews made possible in part by funding from the Great Lakes Center for Education Research and Practice. It is also available at http://greatlakescenter.org.

Review of *The Effects of Test-Based Retention on Student Outcomes Over Time: Regression Discontinuity Evidence from Florida*

Joseph P. Robinson-Cimpian, University of Illinois Urbana-Champaign

I. Introduction

A number of jurisdictions have laws requiring the use of standardized achievement scores in promotion/retention decisions. In general, these laws and policies set passing scores or cut-off scores in order to be automatically considered eligible for promotion to the next grade level. In this working paper,[1] the authors use longitudinal data from students in grades 3 to 12 in all public schools in Florida from school years 2003-4 to 2012-13 to study the short- and long-term effects of third-grade retention on student outcomes.

The report uses regression discontinuity methods, which compare students that score very close together but fall on opposite sides of the cut-off score. This can create a real-world scenario that is similar to a true experimental design. The technique is powerful and is aimed at making causal claims. Studies that use such robust methods could influence early-grade test-based promotion/retention policies across the country, not just in Florida.

II. Findings and Conclusions of the Report

Contrary to the conventional wisdom on grade retention, the report finds that students in Florida who are retained in third-grade because they just barely failed to attain the state's threshold performed better than students of the same age on next year's tests of math and reading. The study does, however, find that these purported benefits fade over time.

III. The Report's Rationale for Its Findings and Conclusions

The report illustrates how less sophisticated methods (e.g., a standard regression approach) can lead to underestimates of the retention effects, due to not accounting for factors unobserved by researchers.[2] Typically, random assignment to experimental and control groups resolves this problem. But this is not always possible on a large scale. Thus, comparing students who barely passed with those who barely failed, and who did not vary from each other on key variables, allows comparisons that would not otherwise be possible. However, this technique requires plausibly satisfying several assumptions. These issues are discussed in Section V below.

IV. The Report's Use of Research Literature

The report cites much of the most directly comparable and appropriate literature. As the authors note in discussing their literature and their new study, their findings are at odds with the conventional wisdom represented in the seminal meta-analysis by Thomas Holmes,[3] who found that retained students performed worse on subsequent tests and were more likely to drop out. One very important distinction between the current report and the individual studies referenced in the meta-analysis is the methodology. The studies in Holmes's analysis compared students across a range of prior achievement levels, and some of the works statistically adjusted for achievement (and other) differences between retained and promoted students. The current report, however, studies the effects of retention at a very narrow window of prior achievement. This is one reason why it may not be wholly accurate to compare the findings of prior research with those of the current report.

The results of this report are, however, consistent with the limited number of recent quasi-experimental studies focused on third-grade retention. These studies include a Chicago study (using an instrumental variables approach) by Jacob and Lefgren and cited in the report,[4] which found no effect on graduation, and a more recent Texas study (using propensity score matching) by Lorence[5] (not cited in the report), which found retained students performed better than promoted students on same-grade tests.

V. Review of the Report's Methods

A strong method for making causal inferences

The report relies on what is known as a regression discontinuity design (RDD), a technique used for making causal inferences from non-experimental data when a threshold determines or strongly predicts treatment assignment. In the case here, all students have a score on the third-grade Florida Comprehensive Achievement Test (FCAT), but only those with scores below the state-specified threshold will be flagged for possible retention according to the policy. Thus, the policy creates a distinction between two sets of students—those who just barely failed to attain the threshold, and those who just barely attained the threshold—that we can otherwise consider to be virtually identical in all respects, except that one set just barely attained the threshold and might now be retained as a result. In effect, the RDD compares the achievement of these two sets of otherwise identical students on later outcomes, such as next-year reading achievement, to estimate the effect of threshold-induced retention flagging in third grade on that outcome.

As the report points out, however, other factors affect retention decisions (discussed in greater detail in the next subsection), and therefore, not all students who fail to attain the threshold are retained. In fact, only about one-third of students who just barely failed

to meet the threshold are retained in grade, compared to about 5% of students who just barely met it. Thus, the authors used a technique known as instrumental variable (IV) estimation to scale the RDD-based estimate of retention eligibility by the proportion of students actually retained as a direct result of the threshold. The IV approach is often used if attaining the threshold *predicts* but does not determine treatment (in this case, retention). For instance, if only half of the people who were assigned to the treatment received it, then we would want to adjust the difference in outcomes by the proportion affected, because only half of the people who we intended the treatment for actually received it.

In this specific case, the difference in next-year reading score is about 24 scale-scored points, or .065 standard deviations (SDs).[6] But because failing to meet the threshold increases the likelihood of retention by only 28 percentage points, the IV approach divides the 24-point difference in outcomes by 0.28, which is the proportion of intended retained students who were actually retained. The logic behind this IV adjustment is that the original 24-point difference is assumed to be the average of a larger effect on 28% of the students and a zero effect on 72% of the students (whose retention status did not hinge on whether they attained the threshold). In this way the 24-point RDD estimate is divided by the 28% of affected students to obtain the retention effect of 84 scale score points, or .226 SDs, on the next year's reading test for students whose retention status was affected by the threshold. This approach of combining RDD and IV methods is common; however, note that the required IV assumption of a zero effect is likely not satisfied, as I discuss in the next subsection, calling the validity of the results into question.

The report provides a number of alternative models and checks to assess the stability of its results. These include the standard RDD checks, such as ensuring that roughly the same numbers of students fall just above and just below the threshold example (see Figure 5 in the report) and that the characteristics of students (e.g., gender, lunch status, race) are similar just above and just below the threshold (see Figure 6 in the report). The report goes further, though, and presents a compelling analysis demonstrating that the students affected by the policy (i.e., the compliers in the IV literature terminology) appear similar on demographic factors to students whose retention status is not affected by threshold attainment (see Table A-5 in the report). Also, the report takes advantage of having comparable data before the policy was implemented (i.e., before 2003) to examine whether other changes at the threshold—specifically, a change in label from "level 2 reader" to "level 1 reader"—could account for the effects. The report finds that this change in label has virtually no effect on outcomes, and thus the observed effects in the main analysis are likely attributable to policy interventions around early-grade retention.[7]

Multiple interventions and intervention groups created by the policy: Serious threats to the report's claims

The report discusses its findings as referring to the "effects of retention," as if the policy creates two sets of students: (1) those who are retained and (2) those who are promoted. However, the policy[8] is much more complex, instead creating *four* sets of stu-

dents depending on whether they were retained or promoted and whether their scores on the third-grade FCAT were above or below the policy threshold: (1) those who are retained and receive additional service, including summer reading camp, (2) those who are retained and receive no other services, (3) those who are promoted and receive additional services, and (4) those who are promoted and do not receive additional services. Table 1 in this review illustrates the services each set of students receives and also provides the percentage of students near the threshold that fall into each category.

Table 1. Services provided to students according to the policy, based on whether the student was retained/promoted and below/above the policy threshold

	Just below threshold	**Just above threshold**
If student is retained…	Group 1: Paragraph (7)(b)(1) of the policy stipulates that, *in addition to* retention itself, students retained under this policy shall receive "intensive instructional services and supports to remediate the identified areas of reading deficiency, including participation in the school district's summer reading camp as required under paragraph (a) *and* a minimum of 90 minutes of daily, uninterrupted, scientifically research-based reading instruction" *and* other services that may include	Group 2: The policy does not specify services for students retained who attained the threshold.
	About 33% of those just below threshold	*About 5% of those just above threshold*
If student is promoted…	Group 3: Paragraph (6)(b) of policy stipulates that students promoted with a "good cause exemption" shall receive "intensive reading instruction and intervention that include specialized diagnostic information and specific reading strategies to meet the needs of each student so promoted"	Group 4: No specified additional services for this group.
	About 67% of those just below threshold	*About 95% of those just above threshold*

As shown in Table 1, according to the policy text, students who are retained because their scores fell below the policy threshold (i.e., Group 1 in Table 1) are not only retained, but also receive "intensive instructional services" including a summer reading camp and at least 90 minutes of daily, uninterrupted reading instruction. By contrast, for the 5% of students just above the threshold who are retained (i.e., Group 2), the policy does not require that the district provide them with any specific services. If the students in Group 2 received the same summer reading camp and intensive instruction as students in Group 1, then a comparison of these two

groups should note that any conclusions refer to the "effects of retention plus summer reading camp and intensive instructional services" rather than simply the "effects of retention." If, however, the services for Groups 1 and 2 truly are different, then the interpretation is not straightforward. Note, however, that the report cannot directly compare these two groups.

In the lower portion of Table 1, we see that the policy also creates two sets of *promoted* students. The Florida policy permits students falling short of the test-based threshold to be promoted provided they have a "good cause exemption," which includes seven categories, such as being an English learner with fewer than two years of ESL instruction or having an IEP that invalidates the FCAT score. According to the policy text, falling just shy of the test-based threshold and being promoted indicates that an exemption was granted and that "intensive reading instruction and intervention" are provided to these students. In these cases (roughly two-thirds of students just below the threshold; Group 3 in Table 1), these services (but *not* retention, since they were promoted) can have a positive effect on subsequent achievement.

Importantly, the use of the IV in this case is inappropriate because it violates the assumption that the threshold has no direct effect on outcomes other than through retention—because it can have a direct effect through the intervention services for promoted students with "good cause exemptions" (the vast majority of cases; 67%). That is, for the report's claim about the causal effect of third-grade retention to be valid, there can be *no other services* that change at the policy-threshold. But paragraph (6)(b) of the policy stipulates that other services change at the threshold—specifically, among all students who are promoted, those who fell short of the cut-off are provided with an intervention that is not specifically provided to those above the threshold. (This concern is in addition to the one raised above about how the policy is not simply a retention policy, but rather a retention-plus-other-services policy.)

The effects of third-grade retention are confounded with the effects of the intensive intervention.

The degree to which this threat is serious or even fatal to the report's analysis depends on whether the policy is actually being carried out. While the policy states a mandate (the district "shall" provide the additional assistance), the report includes no evidence, one way or the other, about whether this provision is being enforced and whether the assistance is actually being provided to help these students. If no assistance is actually being provided, then the report's analyses are not really threatened by the "good cause exemption" issue. If, however, the assistance is substantial, then this issue invalidates the use of the IV and renders the main conclusions of the report correspondingly invalid.

To summarize this subsection, the policy does more than require retention for students below the threshold. Rather, it requires that those retained students receive additional services, and thus, the effects of the report are—at a minimum—a combination of retention effects and effects of the other interventions for retained students (e.g., summer reading camp). In addition to this complication, the policy allows students below the threshold to be promoted, but it requires that these students receive intensive reading instruction and interventions intended to raise their achievement. This additional requirement in the policy text implies

that the RDD+IV assumptions are likely not satisfied, and the main estimates of the report would thus be uninterpretable. Instead, we could perhaps interpret the RDD-based estimates (which are less than 1/3 the estimated size of the RDD+IV estimates) as reflecting a weighted average of the effect of retention, summer reading camp, intensive instruction and additional services for about 28%[9] of students and the effect of the intensive reading instruction and intervention for about 67% of students. In general, though, the report cannot disentangle the effects of retention from those of the summer reading camp, the intensive reading instruction, or any of the other services allowed by the policy (e.g., reduced teacher-student ratios; mentoring or tutoring; more frequent progress monitoring; extended school day, week, or year).

Limits to generalizability

Although the RDD method has tremendous potential for allowing researchers to make causal inferences in the absence of an experiment, the population to which those inferences can be easily generalized is extremely limited. That is, if we set aside for the moment the "good cause exemption" threat to validity, this study may hold lessons about the effects of this retention policy (plus the other services retained students receive) on students at and very near the policy threshold, but we do not know if those effects would be obtained further away from the threshold, say, to students who score half a standard deviation below the threshold. Likewise, we do not know the extent to which students half a standard deviation above the threshold would benefit from this retention policy.

The IV approach (if it is valid) has its own limitations to generalizability, which further restrict the population to which inferences can be made. Namely, we cannot generalize findings to students that teachers (or parents) would definitely want to retain or definitely not want to retain. The former group is small, but the latter seems to be the largest group in the data used for the report, constituting roughly two-thirds of students in the data.[10]

Combining the RDD and IV restrictions to generalizability, the results of the report can only reasonably be generalized to students in Florida retained in third grade who are very near the retention/promotion cusp and whose status hangs on which side of the cusp their third-grade reading test score falls. Looking at Figure 5 of the report, indicating where the threshold lies in relation to the density of observations (i.e., about 1 SD below the mean value), the proportion of the full data at the threshold is approximately 1%, and of that 1%, only 28% are directly affected by the policy. A quarter of one percent is not a very large segment of the population. Therefore, if the application of RDD+IV produced valid estimated effects for retention alone (which is questionable given the "good cause exemption" interventions), we would still be remiss to think that the estimate for this very small subpopulation should generalize to broad retention policies.

Short- vs. long-term outcomes: Some complications

Focusing on the interpretation of the effect estimates, several additional long-term con-

siderations come into play. Regarding the short-term outcomes (e.g., test scores one year after the third-grade retention decision), the report's use of the RDD and IV methods may estimate the causal effect of retention on the subpopulation discussed above (setting aside the "good cause exemption" concerns already noted); however, as the report looks to longer-term outcomes (e.g., test scores six years after the third-grade retention decision, graduation), the interpretation of the estimate becomes more obscured.

The usual suspect for concern regarding longer-term outcomes in longitudinal studies—differential attrition—is not so much of a concern here. The report does a nice job of assessing the potential for differential attrition based on observable factors. The result of that assessment is that differential attrition is likely to be very small at most and unlikely to account for much (if any) of the effects fading over time.

The bigger concern is more intriguing: being retained in third grade lowers the likelihood of later-grade retention. In other words, third-grade retention itself reduces a student's chances of being retained in fourth, fifth, sixth, and seventh grades (see Table 6). Thus, the effect estimate two years after the third-grade retention decision is not simply answering the question "What is the effect of third grade retention on student achievement when students should be in grade 5?", but rather is answering "What is the effect of third grade retention (and other services) *and a lower chance of fourth grade retention* on student achievement when students should be in grade 5?" Note that the report acknowledges this limitation somewhat and provides an estimate of the corrected effect, after accounting for the differential in the proportion of students retained the next year. Bringing this limitation to the fore is commendable; however, the method for correcting the estimate is more assumption-laden and less rigorous than the original estimate.

More specifically, to correct for this combined effect, the report assumes that retention in fourth grade has a similar effect as retention in third grade. Without any evidence in the context of Florida's policy to support this assumption, its validity is uncertain at best and, based on prior studies by Jacob and Lefgren[11] as well as Manacorda,[12] is likely to be incorrect. That is, it appears that later-grade retention often leads to more negative outcomes for graduation. By not properly accounting for these more negative effects, the report may understate the long term-benefits of early-grade retention (and, more accurately, also of additional services for retained students, as well as the intensive reading intervention for students who are promoted due to the "good cause exemptions"). Using their adjustment approach, the authors conclude that differential retention in fourth grade accounts for up to 33% of the decay in the effect on reading and up to 21% in math. First, however, it may be inaccurate to claim this account for "no more than X%" because we do not know if the assumptions are valid, and they appear to contradict both the literature and the report's own assertions earlier that would suggest the later-grade retention effects are less positive. Second, it is likely inappropriate to apply the third-grade effect estimate to later years because, as stated above, the policy indicates that retention in third grade includes other services—services that may not be provided if retained in later grades. Third, later-grade retention effects can have a compounding effect (i.e., students can be retained in several grades and each of them can have lasting effects), but the report does not address this issue.

Another important point is that although the report uses data pooled across six cohorts to estimate the effect one year after the retention decision, the analyses six years out only include data from the oldest cohort. This means that the population of inferences is changing from year to year across the analyses, the sample size is shrinking, and the estimates have additional noise.[13] These limitations grow more substantial as the report estimates longer-term outcomes, and importantly, are in addition to the above concerns regarding the long-term outcome analyses. Finally, the analyses examining retention effects on graduation are also only based on the first cohort affected by the policy, not the six cohorts used for the short-term outcome analysis.

VI. Review of the Validity of the Findings and Conclusions

The report purports to answer the question, "What is the effect of third grade retention on student achievement [when students making normal progress should be] in grade *g* [e.g., 4, 5, etc.]?" (p. 8). The report uses a strong design to estimate the effect of third-grade retention (and other services, such as a summer reading camp) in Florida based on a state-level policy requiring passage of a test-based threshold. That said, the claimed causal effects are questionable because of services provided to students who are promoted despite failing to attain the threshold. Beyond this substantial concern, there are other limitations, such as the population to which the effects can easily be generalized is well less than 1% of the population represented by the data, and the interpretation of the report's estimates become murkier as one moves further away from the time of the retention decision.

VII. Usefulness of the Report for Guidance of Policy and Practice

Because the Florida retention policy stipulates that services must be provided to students promoted with a "good cause exemption," the effects of third-grade retention are confounded with the effects of the intensive intervention. This calls into question both the estimated effects and the utility of this report. Moreover, even if the concern of confounding effects can be overcome, there are other factors that limit the utility of the report. First, the estimated effects do not simply reflect retention alone, but rather retention plus other services. Second, the generalizability of the effects is very limited, applying only to Florida students at or very near the statewide third-grade test threshold for retention and simultaneously directly affected by the policy. Third, differences in later grade retention between those retained in third grade and those not retained, combined with a sample that begins with six cohorts and decreases to a single one, render the longer-term outcome analysis less compelling. Thus, the primary strength of this report may lie in its analysis of short-term effects, but even these effects must be interpreted as valid for only the students very near to Florida's test-based threshold in third grade and are very likely not valid due to confounding intervention effects.

Notes and References

1 Schwerdt, G., West, M.R, & Winters, M.A. (2015). *The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida*. Cambridge, MA: National Bureau of Economic Research.

2 While their argument is clear, the empirical illustration might have been improved by using interaction terms in the regression to estimate the retention achievement differential at the policy threshold. Such an analysis would provide a more direct comparison to the estimate based on the regression discontinuity design, which estimates effects at the policy threshold.

3 Holmes, C.T. (1989). Grade level retention effects: A meta-analysis of research studies. In L.A. Shepard & M.L. Smith (Eds.), *Flunking grades: Research and policies on retention*, (pp.16-33). Philadelphia, PA: The Falmer Press, Taylor & Francis, Inc.

4 Jacob, B.A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, *1*(3), 33-58.

5 Lorence, J. (2014). Third-grade retention and reading achievement in Texas: A nine year panel study. *Social Science Research*, *48*, 1-19.

6 This figure of .065 SDs is based on my calculations taking the year-after estimate from Table 4, column 4, and multiplying it by the baseline effect estimate of threshold attainment on retention (of .283) from Table 6, column 1. The report authors provide the standard deviation of 370 points on the test (from Table 4), which are used to create the standardized effect estimates.

7 The report estimates the label effects separately for cohorts in 2001 and 2002. It is possible, however, to include these early cohorts in a larger analysis with the cohorts affected by the retention policy and subtract off the intent-to-treat effect of cohorts 2001 and 2002 from that of cohorts 2003 and later, then scale the new difference by the IV estimator for the 2003 cohorts and later. My colleague and I used this approach in Robinson-Cimpian, J. P. & Thompson, K. D. (in press). The effects of changing test-based policies for reclassifying English learners. *Journal of Policy Analysis and Management*. See Equation 5 in that paper.

8 For the full policy, see http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=1000-1099/1008/Sections/1008.25.html

9 This 28% comes from the 33% just below the threshold who are retained minus the 5% just above the threshold who are retained. Because of this subtraction, the 28% and 67% only sum to 95% instead of the expected 100%.

10 The report estimates that the retention outcomes of about 28% of students hinge on whether they just barely attained or just barely failed to attain the threshold. Based on the other estimates in Figures 2 and 4, it stands to reason that the roughly 5% of students just barely above the threshold would be retained regardless. Conversely, the two-thirds of students just below the threshold (i.e., retention eligible based on the policy) would not be retained regardless.

11 Jacob, B.A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, *1*(3), 33-58.

12 Manacorda, M. (2012). The cost of grade retention. *Review of Economics and Statistics*, *94*(2), 596-606.

13 Note that these sample changes can create issues that are distinctly separate from the issue of differential attrition, which is not much of a concern in this report, as discussed in the main text.

Document Reviewed:	The Effects of Test-Based Retention on Student Outcomes Over Time: Regression Discontinuity Evidence from Florida
Authors:	Guido Schwerdt, Martin R. West, & Marcus A. Winters
Publisher/Think Tank:	National Bureau of Economic Research (NBER)
Document Release Date:	August 2015
Review Date:	December 3, 2015
Reviewer:	Joseph P. Robinson-Cimpian, University of Illinois at Urbana-Champaign
E-Mail Address:	jpr@illinois.edu
Phone Number:	(217) 333-8527

Suggested Citation:

Robinson-Cimpian, J.P. (2015). *Review of "The Effects of Test-Based Retention on Student Outcomes Over Time: Regression Discontinuity Evidence from Florida."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-NBER-retention.