GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE

DOCUMENT REVIEWED:	"**Getting Farther Ahead by Staying Behind: A Second-Year Evaluation of Florida's Policy to end Social Promotion**"
AUTHORS:	Jay Greene and Marcus Winters
PUBLISHER/THINK TANK:	Manhattan Institute
DOCUMENT RELEASE DATE:	September 13, 2006
REVIEW DATE:	October 10, 2006
REVIEWER:	Derek Briggs
E-MAIL ADDRESS:	derek.briggs@colorado.edu
PHONE NUMBER:	(303) 492-6320

Summary of Review

In the report "Getting Farther Ahead by Staying Behind: A Second-Year Evaluation of Florida's Policy to End Social Promotion,"[1] Jay Greene and Marcus Winters conclude that Florida's recently instituted policy of test-based retention has helped academically struggling elementary school students improve their reading. The evidence provided to support this conclusion is based upon a methodological approach known as an instrumental variable regression analysis. The authors find a small effect for Florida's retention policy on the 2002-2003 third grade cohort after one year, and a more substantial effect for the policy two years after the retention.

While the findings by Greene and Winters are suggestive and merit further investigation, the validity of their conclusions is threatened by the following factors:

1. Florida's retention policy has three major elements; it includes more than just repeating the same grade twice. Retained students are also required to attend a summer school intervention and to receive ongoing intensive reading instruction. The effects estimated by Greene and Winters include all of these experiences. This makes it impossible to isolate the effect of repeating the same grade from the effect of attending the summer intervention and of receiving intensive reading instruction.

2. While the study's methodological approach is in general appropriate for the analysis Greene and Winters have conducted, the authors omit important information necessary to understand and evaluate the particular model they specified. Particularly problematic is the omission of key descriptive statistics about the characteristics of samples analyzed over the study's two-year time period.

Even under the assumption that their instrumental variable regression analysis has been appropriately specified, the authors appear to misinterpret the retention effect they have estimated. The upshot of this misinterpretation is that the magnitude of both the one- and two-year retention effects are overstated.

Review

I. INTRODUCTION

The state of Florida recently instituted a policy of grade retention, attempting to end the practice of social promotion. The policy, enacted as of the 2002-2003 school year, mandates the retention of third grade students scoring below a set threshold on the state's standardized achievement test for reading, the Florida Comprehensive Assessment Test (FCAT). Prior to the policy, decisions about promotion and retention were left to the discretion of school district personnel. At issue is the effectiveness of this new policy with respect to changes in the reading achievement of retained students.

The 2006 report "Getting Farther Ahead by Staying Behind: A Second-Year Evaluation of Florida's Policy to End Social Promotion" by Jay Greene and Marcus Winters[2] is a follow-up to a 2004 report by the same two authors entitled "Getting Ahead by Staying Behind: An Evaluation of Florida's Policy to End Social Promotion".[3] The latter report, in which the authors estimated a positive effect for Florida's policy, was previously reviewed by Wiley,[4] who was quite critical of that report's methods and findings. While some of Wiley's criticisms of the first report are also applicable to the second (as discussed below), in this review I focus attention upon the way Greene and Winters have estimated the one-and two-year effects of Florida's retention policy, and the reasons why these effect estimates may have an equivocal interpretation.[5]

II. REPORT'S FINDINGS AND CONCLUSIONS

Greene and Winters find that students who have been retained under Florida's policy outperform comparable groups of students who were promoted. The policy appears to be more effective for retained students after two years than it is after one. Using what the authors consider their most generalizable set of analyses, the estimated effect of the policy on FCAT reading scores after one year is about 4 points.[6] After two years, the

effect for the same students is about 41 points. When reported in effect size units as a percentage of a standard deviation in FCAT scores, the magnitude of the one-year effect (.01) is interpreted by the authors as "small,"(p. 9) while the magnitude of the two-year effect (.11) is interpreted as "moderate" (p. 10).

The authors conclude that Florida's test-based retention policy is "helping students improve their reading" (p. 12). They are careful not to generalize their findings to other states and cities instituting similar policies. While Greene and Winters express uncertainty about whether these benefits are likely to "continue to hold, expand or disappear over time" (p.11); and about whether the benefits of the policy outweigh the costs,[7] they conclude that "any large-scale policy that produces progress is promising" (p. 12).

III. REPORT'S RATIONALES FOR ITS FINDINGS AND CONCLUSIONS

Greene and Winters have conducted a quasi-experiment to estimate the effects of Florida's test-based retention policy over a two year period. The treatment group consists of the third grade students who scored below the promotion threshold (1,045) on the FCAT reading test at the end of the 2002-2003 academic school year, and were thus subject to the retention policy. The scores of these students one and two years later are then evaluated relative to two different control groups, each of which is intended to represent the change in FCAT reading scores that would have been observed had retained students been promoted.

Control Group #1

The cohort of third grade students during the 2001-2002 academic school year who scored below the promotion threshold on the FCAT, but who were not subject to Florida's test-based retention policy. This control group is the basis for what the authors refer to as their "across-year comparison."

Control Group #2

A subset of third grade students within the same 2002-2003 cohort who scored just barely above the FCAT threshold for promotion. This control group is the basis for what the authors refer to as their "regression-discontinuity comparison."

While control group #1 is being compared with all students in the 2002-2003 cohort who were subject to the new retention policy, control group #2 is being compared with a subset of the 2002-2003 cohort that scored just barely below the promotion threshold. The authors quantify "just barely" two ways: 25 and 50 points away from the score threshold.

The one- and two-year effects of the retention policy are estimated using what is known as an instrumental variable regression analysis. This is what the authors refer to as "two-stage approach" (p. 8). For a description of this methodological approach as it has been implemented by Green and Winters, please see the technical appendix to this review.

The instrumental variables regression analysis is an entirely defensible approach for the kind of data that has been gathered. The reader is being expected to take much of this as a matter of faith, however, because the authors provide no information as to the details of their model specification in their report.

IV. REVIEW OF THE REPORT'S USE OF RESEARCH LITERATURE

In reviewing prior research concerning the effects of retention on academic achievement, Greene and Winters make the distinction between studies of "discretionary retention" and "test-based retention." In the former class of studies, the decision to retain a student is left to the discretion of educators; in the latter class, the decision is based upon performance on a standardized test. The authors are interested in evaluating the effectiveness of the latter as a policy approach, and they note that few such studies have been conducted because the use of test-based retention policies is a relatively recent phenomenon. The authors' first year report contained a more extensive review of the literature on discretionary retention studies. In the current report, their review of previous research on test-based retention is brief and focuses exclusively upon a 2005 study published by Roderick and Nagaoka .[8]

Roderick and Nagaoka evaluated the effects of a test-based retention policy implemented for the Chicago Public School system between 1997 and 2000. In general, they found roughly no effect for the policy after one year, and a negative effect for the policy after two years. Greene and Winters are not entirely persuasive in their attempts to reconcile their findings of a moderate effect in the second year of Florida's retention policy with the moderate negative effect found in Chicago by Roderick and Nagaoka. They attribute the contradictory findings to differences between the two policies and their implementation, however. It is worth noting that Greene and Winters have very much changed their tone regarding the Roderick and Nagaoka study since the time of their previous report. Whereas before they were quite critical of what they describe as Roderick and Nagaoka's regression-continuity comparison, they have now adopted the same approach for their own study.

There are at least two important omissions from the authors' review of the literature. First, they make no reference to a study—published in the very same issue of the journal in which the Roderick and Nagaoka study was published—by Allensworth in which she estimates the effect of Chicago's retention policy on subsequent dropout rates following the eighth grade of school.[9] Allensworth finds that the policy was associated with an increase in dropout rates. This finding is important to present because it suggests the possibility that even if a test-based retention policy can be shown to have positive effects on short-term academic achievement, it might at the same time be causing students to drop out of school. The authors also overlooked a second recent review, entitled "Retention, Social Promotion, and Academic Redshirting: What Do We Know and Need to Know," recently published in the journal *Remedial and Special Education*.[10] This rather comprehensive review describes many important nuances that would useful information for any city or state contemplating the sort of test-based retention policy currently mandated in Florida.

V. REVIEW OF REPORT'S METHODS

There are three significant problems with the methodological approach taken by Greene and Winters: (1) the omission of important descriptive statistics, (2) the timing of student testing in the baseline year, and (3) the interpretation of their instrumental variable regression analysis.

Omission of Descriptive Statistics

The authors present some useful information in their Tables 2-3, which contain baseline year characteristics of their treatment and

control groups used for both the across-year and regression-discontinuity comparisons. Unfortunately, additional descriptive statistics along these lines are not presented for these groups in the two years following their initial retention or promotion. This is especially important with respect to students who were retained in third grade during the 2002-2003 school year. By the 2003-2004 school year, for example, it is natural to wonder how many of these students were promoted to the fourth grade, were retained again, or left the state altogether. These and other descriptive statistics are missing from the report. Nor do Greene and Winters explain discrepencies in the descriptive statistics that are provided. For example, the combined sample size of the treatment and control groups presented in Table 2 is 88,565. But the sample size used in their instrumental variable regression equation for the one-year retention effect is only 79,747. What happened to these missing 8,818 students?

Timing of Student Testing

The outcome of interest when estimating the one-year effect of the retention policy is the change in FCAT reading scores from the end of one academic school year to the end of another. Yet for the cohort of students retained after the 2001-2002 school year, these two test administrations encompass not just their second year spent in the third grade, but their experience in a mandatory summer school intervention that predates their grade retention. In contrast, the baseline testing in the retention study conducted by Roderick and Nagaoka took place after "to be retained" students participated in a similar summer school intervention.[11] This was done to isolate the effect of retention relative to the effect of the summer school intervention. Because the one-year retention effect estimated by Greene and Winters combines both summer school and retention, it is entirely possible that the one-year effect of summer school is largely positive while the effect of retention is largely negative, or vice-versa. It is not possible for Greene and Winters to isolate the effect of simply repeating the same grade in their analysis.

Interpretation of Instrumental Variable Regression Analysis

There are three potential problems with the way the authors have presented the key results from their instrumental variable regression analysis. First, it can be shown (see the Technical Appendix to this review) that the results have been presented in a way that is very likely to overstate the effects of Florida's retention. Second, the authors include no models in which statistical interaction terms have been included. This makes it impossible, for example, to examine whether the effect of retention is bigger or smaller as a function of a student's race/ethnicity or free-lunch status. Third, the authors do not report all parameter estimates for the full set of variables included in their analysis.[12] That these estimates are sensible must be taken on faith. It is important to provide the full set of parameter estimates from the regression analysis (at least as part of an appendix to the report), because if some of these estimates appear counterintuitive, it raises questions about the way the underlying statistical model has been specified. Again, for more on this issue, see the Technical Appendix.

VI. REVIEW OF THE VALIDITY OF THE FINDINGS AND CONCLUSIONS

There appear to be three significant threats to the validity of the causal inference advanced by Greene and Winters that Florida's retention policy is helping students improve their reading.

1. The policy studied is not simply retention.
2. Possible misspecification of the instrumental variable regression model.
3. Misinterpretation of the magnitude of the estimated policy effect.

The Policy Studied is Not Simply Retention

The analysis reported by Greene and Winters should not properly be understood as evaluating what people are apt to think of when they hear the term "retention. " Grade retention is typically assumed to mean that a student simply repeats the same grade in school. This is not all that happens under Florida's policy. A third grade student scoring below the FCAT threshold for promotion is immediately expected to develop, in consultation with educators and parents, an academic improvement plan. As part of this plan the student is to be provided "intensive reading instruction." This intensive instruction begins as part of a summer school intervention and continues when the student repeats the third grade in the subsequent school year. Indeed, Florida's policy is intended to ensure that retention does *not* mean simply repeating the same educational experiences twice, and this is what constitutes both the one-year and two-year treatment that Greene and Winters are evaluating.

A different problem is that by year 2 of the study, there are two different retention experiences combined into a single effect estimate.

- The experience of those who are retained for a year (which includes summer school and ongoing intensive reading instruction) and then promoted into the fourth[h] grade, and
- The experience of those who are retained twice in the third grade (again including summer school and ongoing intensive reading instruction).

It appears that the latter is a relatively small proportion of the treatment sample.[13] However, if those who are retained twice benefit more from retention than their promoted counterparts, it is likely to bias the two-year effect of "retention" somewhat upwards. On the other hand, if those students retained twice are ones who benefit the least from retention, it is likely to bias the two-year effect somewhat downwards.

Possible Misspecification of the Instrumental Variables Regression Model

The authors provide no information that would allow the reader to assess the sensitivity of the instrumental variable regression they have specified. In contrast, Roderick and Nagaoka specify both a traditional multiple regression, followed by an instrumental variable regression. [14] This allows the reader to assess the extent to which the estimated retention effect found in the instrumental variables regression differs from that found in the traditional multiple regression. This is especially helpful in illuminating the direction of selection bias—does applying the instrumental variable regression relative to a traditional multiple regression result in a smaller estimated effect or a larger one? How much bigger or smaller? If the effect changes from large and negative to large and positive, or otherwise changes dramatically in magnitude, it might lead us to question the proper specification of the statistical adjustment.

A related problem is that the authors have chosen to specify a model that includes no statistical interactions. Does retention have the same effect on Hispanic students as it does on White students? Does it have the

same effect on students in high-poverty school districts relative to students in low poverty school districts? These more nuanced questions about Florida's retention policy are important to ask, as it might change the subsequent focus of the policy if retention has negative effects for certain subgroups of students, but positive effects for others.

Misinterpretation of the Estimated Policy Effect

There are two ways in which the authors appear to be misinterpreting the size of the estimated policy effect. The first is presented in the Technical Appendix and involves a potential misinterpretation of their model that results in an overstatement of the magnitude of the estimated retention effect.. The second is in their characterization of the effect sizes for the one- and two-year effects as "small" and "moderate." Whether the estimated effects found here are presented in standard deviation units or percentile units (see p. 10 of the report), from the standpoint of practical significance I would characterize the one-year effects found by Greene and Winters as insubstantial, and the two-year effects as small. As the authors note, most retained students are scoring in the baseline year at only about the 23rd percentile of the FCAT. Even if—under a very optimistic scenario—retention were to help move these students up 5 percentile points after two years (as the authors suggest on p. 10), it seems unlikely that this level of performance would suffice for them to be classified as proficient in reading anywhere in the near or distant future under the federal stipulations of the No Child Left Behind policy.

VII. REPORT'S USEFULNESS FOR GUIDANCE OF POLICY AND PRACTICE

The findings from this report are best described as suggestive. They indicate that there may be a small positive effect associated with Florida's test-based policy for students in their second year after being retained. It remains unclear why little to no effect in the first year is followed with a small effect in the second year. This finding requires further investigation so that it can be better explained and understood. It is possible that the effects estimated for retention in this report are artifacts of misspecified statistical models. This also merits further investigation.

One useful aspect of the report is its attempt to contrast the retention policies of Florida with those of Chicago. The authors acknowledge that these comparisons are largely speculative, however. The report provides little guidance as to the ideal practice of implementing the retention policy with students, which is probably the issue that should be of most interest to policy makers. Greene and Winters appear to believe that students who are struggling academically will get "farther ahead" by "staying behind" because they have been retained. If the Florida policy of combining summer school, intensive reading instruction, and grade retention is shown to have positive effects on students' subsequent FCAT scores, however, policy makers will want to know more about each of these separate interventions. Might the intensive reading instruction be carrying the load all by itself? In fact, might summer school and grade retention each have a negative effect? Similar scenarios could be envisioned for each of the interventions. Or perhaps there is a combined effect that enhances each independent element. These are questions that should be explored in subsequent research.

TECHNICAL APPENDIX: Instrumental Variables Regression Analysis

Greene and Winters use an instrumental variable regression analysis to estimate the effects of Florida's test-based retention policy. To most readers perusing the findings of their report, however, it might appear that the authors have estimated the effects of Florida's retention policy using a traditional application of a linear regression equation along the lines of

$$Test\ Score\ Change = a + b * Z + c_1 * X_1 + c_2 * X_2 + ... + c_P * X_P . \quad (1)$$

In the regression equation above, the outcome variable on the left side of the equation, "Test Score Change," represents the predicted change in FCAT score for a student from the baseline year to the first or second year that follow. This predicted change has been modeled as a function of the variables on the right hand side of the equation. The variable *Z* will equal 1 if the student was retained and 0 if the student was promoted. The variables $X_1, X_2, ..., X_P$ (where the subscript *P* is used to represent the total number of variables included) represent the observed characteristics of the students that might confound the relationship between retention status and test score change. These are often referred to as "control" variables. As specified by Greene and Winters, these include baseline FCAT score, racial/ethnic status, free and reduced-price lunch status, limited English proficiency, and the school district with which a student was affiliated.

From a policy standpoint, the key parameter estimate of interest in this regression equation is *b*, and it has both a technical and a substantive interpretation. The technical interpretation of *b* is that it represents the marginal test score change associated with a single unit increase in the variable *Z*, holding constant the values of all other variables. Because a single unit change for *Z* represents the difference between being in the control condition (i.e., promoted, $Z = 0$) and the treatment condition (i.e., retained, $Z = 1$), *b* can be interpreted substantively as the effect of Florida's test-based retention policy. So if $b = 4$, the average policy effect would be 4 points on the FCAT, if $b = 40$, the average policy effect would be 40 points, and so on.

Greene and Winters, however, are using a slightly different version of the approach described above. The reason for this is that not all students in the treatment group are ultimately retained, and not all students in the control groups are ultimately promoted. For example, the authors point out that 43 percent of students in the 2002-2003 cohort that scored below the FCAT promotion threshold were nevertheless able to obtain a waiver exempting them from the state policy. If only those students in the treatment group that were actually retained were compared with those students in the control group that were actually promoted, there is a good chance that the estimated effect of the retention policy would be biased. For example, students who are able to obtain exemptions might on average perform better (or worse) on the FCAT than their counterparts unable to obtain an exemption. This

problem, known as selection bias, is a problem typically associated with quasi-experimental study designs.

Noting this problem, Greene and Winters briefly allude to the approach they take to address it.

> When there are a lot of exemptions, we risk running into the same methodological dangers that beset earlier studies of discretion-based retention. If exemptions are granted on a discretionary basis, perhaps retained students will once again be incomparable in key unobserved ways. To address this problem, we use a two-stage model. In a two-stage approach, we essentially identify who would have been retained if exemptions did not distort the pool of retained students. Then we predict the effect of this undistorted retention on academic achievement. This technique removes bias that could be introduced by the subjective use of exemptions. (p. 8)

The two-stage approach to which the authors refer (but do not elaborate in their Manhattan Institute report[15]) would proceed as follows for the across-year comparison:

> Stage 1. Predict the probability that a given student will be retained ($\hat{Z}$) regardless of whether that student is in the 2001-2002 or 2002-2003 third grade cohort.
>
> Stage 2. Replace the variable Z with $\hat{Z}$ and then estimate the new regression equation

$$Test\ Score\ Change = a + b * \hat{Z} + c_1 * X_1 + c_2 * X_2 + ... + c_P * X_P \qquad (2)$$

The same two-stage approach described above is used for Greene and Winter's regression-discontinuity comparison. The only difference is in the sample of students considered, the way that the probability of retention is predicted in stage 1, and the variables included for $X_1, X_2, ..., X_P$ in stage 2.

Some details about the two stages summarized above are in order. In stage 1, the probability of retention is predicted as a function of many of the same variables included in the regression equation in stage 2 (i.e., baseline reading scores, race/ethnicity, free and reduced lunch status, etc.). There must be at least one variable used in the first stage that is not included in the second, however. This variable, known as an "instrument," must have the following characteristics: (a) it is strongly correlated with actual retention status (i.e., *Z*), and (b) it is uncorrelated with any variables that have been omitted from the regression equation. Note that no matter what variable is chosen to fill this role in stage 1, only characteristic (a) can be demonstrated empirically. That is, characteristic (b) can never be empirically validated because the variables that were omitted from the regression equation

are unobserved. This is one reason why the results from an instrumental variables regression analysis are always equivocal to some extent.

In the across-year analysis by Greene and Winters, the instrument being used is a dummy variable that indicates whether or not a given student was a member of the 2001-02 third grade cohort (before the test-based retention policy took effect) or the 2002-03 third grade cohort (after the test-based retention policy took effect). We might expect this variable to be pretty strongly correlated with actual retention status, but uncorrelated with the unknown reasons why some students have bigger test score changes than others. Greene and Winters indicate that there is a strong correlation between the instrument and actual retention status in the more technical version of their report.[16] One might reasonably question the lack of correlation between the instrument and omitted variables, however. As the authors note in their technical report, Florida has many other policies intended to improve student achievement, and as these policies mature we might expect them to become increasingly effective. This is why the authors also employ their regression-discontinuity comparison.

In the regression-discontinuity comparison, the instrument being used in stage 1 to predict retention status is a dummy variable that indicates whether or not a student has scored above or below the promotion threshold on the FCAT. While this variable has a strong correlation with retention status, however, it seems a bit odd to posit that it will be uncorrelated with omitted variables that help explain why students have high or low test score gains. I point this out as an illustration of one way in which the specification of an instrumental variable regression analysis can be called into question.

Now we return to the regression equation that gets specified in stage 2 of the analysis. Note that the only difference between regression equations (1) and (2) is that Z (retention status), has been replaced by $\hat{Z}$ (probability of being retained). Unlike the variable Z, the variable $\hat{Z}$ does not take on the dichotomous values of 0 or 1 for each student in the sample, but a probability that ranges *between* the values of 0 and 1. These values represent the probability that a given student is actually retained in the third grade.

The parameter estimate for b in equation (2) is what Greene and Winters have presented in their Table 4 (p. 9) in the rows labeled "Across-Year Comparison," "Regression Discontinuity—Within 25 points" and "Regression Discontinuity—Within 50 points. " But should these parameter estimates be interpreted as the effect of Florida's test-based retention policy? The answer is probably not, because the technical interpretation of b as the marginal test score change for a *unit* change in $\hat{Z}$ represents something we do not actually observe in the data. That is, a unit change in $\hat{Z}$ represents a change in the probability of being retained that shifts from 0% to 100%; but the actual values of $\hat{Z}$ for the sample of students being analyzed will be somewhere between these two extremes. To assume that the difference be-

tween students in "treatment" and "control" conditions represents a difference in 100% for $\hat{Z}$ represents a dubious extrapolation.

Even in the technical version of their report, Greene and Winters do not reveal the range of values estimated for $\hat{Z}$ among those who were actually retained and those who were actually promoted. We would expect, on average, the former to be high (i.e., 90% probability of being retained), and the latter to be low (i.e., 20% probability of being retained). While these values are unknown, it is safe to assume that the difference between the two will be something less than 100%. It is this difference that must be multiplied by the estimated effect parameter *b* in the regression equation. For example, in Greene and Winters' across-year comparison in Table 4 they report a two-year effect estimate for the retention policy of about 41 points. If the average difference in $\hat{Z}$ between those retained and those promoted were 70%, then the effect of the retention policy should not be interpreted as 41 points, but as 0.7*41 points = 29 points. Along these lines, if the average difference were less than 0.7, the estimated effect would be smaller; if the difference were greater than 0.7, the estimated effect would be larger.

NOTES & REFERENCES

[1] Greene, J. & Winters, M. (2006, September). *Getting farther ahead by staying behind: a second-year evaluation of Florida's program to end social promotion.* Civic Report. New York, NY: Manhattan Institute.

[2] Greene, J. & Winters, M. (2006, September). *Getting farther ahead by staying behind: a second-year evaluation of Florida's program to end social promotion.* Civic Report. New York, NY: Manhattan Institute.

[3] Greene, J. & Winters, M. (2004, December). *Getting ahead by staying behind: An evaluation of Florida's program to end social promotion.* Education Working Paper. New York, NY: Manhattan Institute.

[4] Wiley, E. (2006, Feb. 23) Review of Greene & Winters, ""An Evaluation of Florida's Program to End Social Promotion," and "Getting Ahead by Staying Behind: An Evaluation of Florida's Program to End Social Promotion". Retrieved Oct. 2, 2006, from http://www.asu.edu/educ/epsl/EPRU/ttreviews/EPSL-0602-119-EPRU.pdf

[5] The ability to explore these issues has been greatly facilitated by the availability of a more technical version of the report that Greene and Winters plan to submit for publication in a peer-reviewed journal. The authors were kind enough to share with me a draft version of that manuscript.

[6] It is worth noting that this effect is roughly 13 points lower than the effect that was presented in the first-year report that Wiley (2006) had reviewed. In the current report Greene and Winters explain this discrepancy as a function of "revisions" to the original data set and the use of "slightly different analytical models" (p. 14). It would have been helpful it the authors were more transparent in this regard, explaining how and why these revisions were made.

[7] According to Greene and Winters, the average cost for a state to retain a single student in grade school for one additional year is about $10,000.

[8] Roderick, M. and Nagaoka, J. (2005) Retention under Chicago's high-stakes testing program: helpful, hurtful, or harmless? *Educational Evaluation and Policy Analysis*, 27(4), 309-340.

[9] Allensworth, E. (2005) Dropout rates after high-stakes testing in elementary school: a study of the contradictory effects of Chicago's efforts to end social promotion. *Educational Evaluation and Policy Analysis*, 27(4),341-364.

[10] Frey, N. (2005) Retention, social promotion, and academic redshirting: what do we know and need to know. *Remedial and Special Education,* 26(6), 332-346.

[11] I thank Melissa Roderick for drawing my attention to this in an e-mail communication.

[12] The full set of parameter estimates is, however, available in the more technical version of their report, available from the authors upon request.

[13] Personal communication from Marcus Winters, September 25, 2006.

[14] Roderick, M. and Nagaoka, J. (2005) Retention under Chicago's high-stakes testing program: helpful, hurtful, or harmless? *Educational Evaluation and Policy Analysis*, 27(4), 309-340.

[15] The authors do provide considerably more detail in the manuscript version of their report.

[16] They do not provide the actual magnitude of this correlation.

The Think Tank Review Project is made possible by funding from the Great Lakes Center for Education Research and Practice.