Summary of Review

In *Getting Ahead by Staying Behind: An Evaluation of Florida's Program to End Social Promotion* (Education Next, February, 2006), Jay Greene and Marcus Winters report the positive effects of Florida's program to end social promotion. This material was originally set forth in a more detailed report published as a Manhattan Institute "working paper" in December, 2004

The authors claim to have found substantial positive effects associated with retention. However, the validity of these claims is questionable due to weaknesses in the analyses on which they are based. Furthermore, the publications omit key information that would allow well-known threats to validity to be addressed in a straightforward way.

This review explains several major flaws of the study, each of which threatens the validity of the study's results and seriously weakens the authors' claims concerning the effectiveness of Florida's retention policy. In summary:

1. Contrary to the authors' claims, the two groups that they compare are not comparable. Florida statewide reading scores reveal that the 2002 (pre-retention policy) cohort started out substantially lower in reading achievement than the 2003 cohort.

2. Incomparability of gains at different parts of the score scale make it impossible to validly interpret the authors' analyses based on gain scores.

3. Regression to the mean, one of the major threats to the validity of research about grade retention, appears to be substantial in this study. The authors' own analyses make this clear, though they make no mention of it.

4. The authors fail to include a key variable in their statistical analysis – the interaction between retention and initial level of achievement. As such, their statistical model rests on the unrealistic assumption that retention helps all students in the exact same way – regardless of whether they are at the bottom of Level 1 or toward the top of Level 2.

5. For each of their statistical models, the authors report only a fraction of the evidence necessary to determine the accuracy of the statistical results. In particular, the regression results do not include intercept estimates that are key to interpreting such results.

Each of the authors' conclusions regarding the impact of retention on achievement is based on their statistical models. Evidence supporting the accuracy of these models is extremely limited and the influence of threats to validity is disturbingly obvious. Accordingly, none of the authors' conclusions can be considered valid in the absence of substantial additional information. The conclusions reported in the two papers are therefore without warrant and should not be relied upon to guide policy or practice.

Review

I. INTRODUCTION
In *Getting Ahead by Staying Behind: An Evaluation of Florida's Program to End Social Promotion* (Education Next, February, 2006), Jay Greene and Marcus Winters report on their study of Florida third-graders held back as part of the state's program to end social promotion (their original study, which is summarized in the *Education Next* article, was entitled *An Evaluation of Florida's Program to End Social Promotion* and was released by the Manhattan Institute in December, 2004). Although substantial research supports the negative effects of grade retention, several state and urban school systems (including New York City and Chicago) have recently implemented policies that rely on grade retention as a basic element of their plans to increase student achievement. The Florida program is prominent among these policies; evaluations of this program will be informative not only in Florida but nationwide.

II. CONCLUSIONS AND FINDINGS

The study reports two major conclusions. First, students subject to retention policies (low-performing third graders in the first year of the policy) outgained students from earlier cohorts, who were not subject to retention policies (low-performing third graders from the year immediately preceding policy implementation). Second, retained students outgained their low-performing counterparts who received exemptions and were promoted to the next grade.

III. BASES FOR CONCLUSIONS AND FINDINGS

The authors compare achievement scores of three groups of third-grade students to estimate the effects of the retention policy and of actual retention. The policy took effect in 2002-2003; therefore cohorts from two successive years – 2001-2002 and 2002-2003 – were used to represent "pre-retention policy" and "retention policy" effects, respectively. These cohorts were compared via multiple linear regression. A second distinction is made within the 2002-2003 cohort. Even though the policy was operating during that year, exemptions were granted to a sizable proportion of third-grade students that year (the exact proportion is unclear; the authors report on page 5 that 21.3% received exemptions but suggest on page 6 that the number was closer to 40%).[1] The authors use this distinction to compare the effects of retention and promotion on the two different subsets of the 2002-2003 cohort. Exemptions were granted on the basis of several criteria, including demonstration of proficiency on either another standardized exam or through a performance portfolio. Granting of exemptions was not random; students granted exemptions were likely different from those who did not receive exemptions. To deal with this incomparability, the authors reportedly used an instrumental variables approach to linear regression; how-

ever, they provide very few details of their analyses.

The analyses are all grounded on two sets of students' assessment scores: on the Florida Comprehensive Assessment Test (FCAT) and the Stanford 9. However, the publications leave open major questions about the samples used. In non-experimental (non-randomized) studies, sampling and potential bias issues are extremely important, and the numbers presented by the authors raise a bright red flag with regard to such sampling. The authors state that they "obtained individual student-level test scores on the math and reading sections of the FCAT and Stanford-9 for the entire population of students in the state of Florida who met the necessary criteria to be part of [their] study."[2] Yet sample sizes reported for FCAT and Stanford 9 analyses vary substantially, with FCAT analyses hovering around 89,000 and Stanford 9 sample sizes reaching a much larger 873,000. The authors offer no explanation for why the Stanford-9 sample reported is nearly ten times the size of the FCAT sample.[3]

IV. REVIEW OF THE REPORTS' USE OF RESEARCH LITERATURE

The authors note that overwhelming evidence supports the contention that students held in a grade for multiple years suffer long-lasting negative academic and emotional effects. They select two prominent studies critical of retention (Holmes, 1989, [4] and Nagoaka & Roderick, 2004[5]) and detail potential shortcomings with each. Yet they claim that the majority of the wide body of research represented by these two studies is "severely limited" due to the nature of the *subjective* retention policies studied. The authors contend that the strength of their own study is the *objective* nature of the Florida retention policy: "The existence of an

objective retention policy in Florida allows for the development of an adequate comparison group not available in previous evaluations." The authors conjecture that the objectivity of a retention policy is more beneficial to students because "…it is possible that the potentially harmful stigma currently associated with retention might not apply to the same extent under the new system, which holds back much larger numbers of students." The authors imply that failure will be widespread under the new policy, but it will be a kinder, gentler, *feel-good failure*. This would indeed be a remarkable outcome of the retention policy. To date, however, neither they nor others have reported results on changes in stigma associated with Florida's grade retention policy.

V. REVIEW OF THE REPORTS' METHODOLOGY

Having established retention policy objectivity as the primary factor distinguishing their study from the majority of previous research on retention, the authors provide the results of several statistical tests that support their ultimate conclusion: "…this study indicates that the use of objective testing to end social promotion leads to substantial academic gains for low-performing students."

This is a strong claim, but it completely disregards many well-known problems of evaluations of retention programs, including lack of randomization of students, regression to the mean, and incomparability of test score gains at different places on the score scale.[6] These and other problems common to retention program evaluations call into question the comparability of any groups that were not randomly assigned (such as those retained and those not retained). For a study such as this to have practical use, it must include a report of information necessary to test whether some of these problems might threaten the validity of conclusions.

Only after verifying that these problems do not affect the results would it be appropriate to conclude that the intervention (retention) accounted for any observed group differences. These authors, though, provide few details behind their analytical method aside from simply identifying the approaches used: linear regression and instrumental variables approaches.

Such selective reporting greatly increases the difficulty of critically reviewing a study. However, several major threats to validity in this study are readily apparent, notwithstanding the information omissions. Described below are five flaws of the study, each of which threatens the validity of the study's results and seriously weakens the warrants for the authors' claims of the effectiveness of Florida's retention policy. In summary:

a. Contrary to the authors' claims, the two groups that they compare are not comparable. Florida statewide reading scores reveal that the 2002 (pre-retention policy) cohort started out substantially lower in reading achievement than the 2003 cohort.

b. Incomparability of gains at different parts of the score scale make it impossible to validly interpret the authors' analyses based on gain scores.

c. Regression to the mean – one of the major threats to the validity of retention research – appears substantial in this study. The authors' own analyses make this clear, though they make no mention of it.

d. The authors fail to include a key variable in their statistical analysis – the interaction between retention and initial level of achievement. As such, their statistical model rests on

the unrealistic assumption that retention helps all students in the exact same way – regardless of whether they are at the bottom of Level 1 or toward the top of Level 2.

e. For each of their statistical models the authors report only a fraction of the evidence necessary to determine the accuracy of the statistical results. The regression results do not include intercept estimates that are key to interpreting such results.

1. *Students subject to the retention policy and students not subject to the policy are not comparable.*

The validity of the authors' analyses rests on the comparability of two groups – low-performing third-graders in 2002 (the year before the retention policy took effect) and their low-performing counterparts from 2003 (the first year of the policy). In the absence of random assignment of students to treatment conditions, policy analyses often must rely on comparing two groups that differ only in whether they are subject to a given policy. The authors claim this is the case for their analysis: "The students from both school years are very similar in all respects except for the year in which they happened to have been born, making comparisons between their improvements particularly meaningful."[7]

The authors could have reported the average scores of each cohort; they did not. They did, however, acknowledge in a follow-up conversation that the groups "were actually kind of different" at the start.[8] In fact, Florida statewide reading scores reveal that the 2002 (pre-retention policy) cohort started out substantially lower in reading achievement than the 2003 cohort. For example, the percentage of Florida 3rd-grade students at

the lowest achievement level was 17% greater in 2002 than in 2003.[9]

The study relies on analyses of gains in FCAT and Stanford 9 scores. It asks (a) how much students gained in achievement from year to year, and (b) whether the average gain in achievement was different for cohorts of retained and promoted students. Analyses based on gain scores can be misleading for a number of reasons. Most significantly, cohorts might start out at different achievement levels, suggesting that their gains are not comparable. Responsible reporting requires that both starting point and gains be provided; without both it is impossible to verify the validity of gain score changes because many potential threats to validity cannot be ruled out. It would suffice to provide a straightforward chart reporting the mean and standard deviation of first-year scores by cohort. Histograms of first-year scores by cohort would be even better, as they would illuminate departures from normality and differences in the cohort that might be masked by considering only means and standard deviations. *The authors fail to provide this straightforward yet critically important information.*

2. *Incomparability of gains at different parts of the score scale make it impossible to validly interpret the authors' analyses based on gain scores.*

Omission of starting point information in this study may be even more problematic than the omission of the means and standard deviations. Gain scores based on FCAT reading can be especially misleading. As detailed in Figure 1,[10] Florida divides reading proficiency into five levels. Level 1 students (students at the lowest of proficiency) are those that, according to the state, have "little success with the challenging content of the Sunshine State Standards."[11] Level 2 students have "limited success" with

those standards. A Level 5 student, in contrast, "has success with the most challenging content…[and] answers most of the test questions correctly, including the most challenging questions." Importantly, the score intervals for each level vary greatly. A Level 1 student can start out with scores in third grade ranging anywhere from 86 to 1045. A student toward the bottom of Level 1 could gain more than 900 points from year to year and still find himself in Level 1. A similar gain from a Level 2 student may propel her into Level 5 – the highest possible proficiency level. Simply put, lower-scoring students have more room to grow, and learning gains at lower levels are represented by greater changes in scores. Clearly, starting point must be considered to interpret any score gains measured by the FCAT. Yet, the authors fail to provide this critical information.

Figure 1

Grade	Level 1	Level 2	Level 3	Level 4	Level 5
3	86-1045	1046-1197	1198-1488	1489-1865	1866-2514
4	295-1314	1315-1455	1456-1689	1690-1964	1965-2638
5	474-1341	1342-1509	1510-1761	1762-2058	2059-2713
6	539-1449	1450-1621	1622-1859	1860-2125	2126-2758
7	671-1541	1542-1714	1715-1944	1945-2180	2181-2767
8	886-1695	1696-1881	1882-2072	2073-2281	2282-2790
9	772-1771	1772-1971	1972-2145	2146-2297	2298-2943
10	844-1851	1852-2067	2068-2218	2219-2310	2311-3008

Chart of FCAT Achievement Levels and FCAT Scores — Reading — On grade-level. Prior to 2002, a three-digit scale score was reported for students.

Florida Department of Education (2004)

3. *The authors fail to consider regression to the mean.*

Students subject to retention policies score at the lowest level on achievement measures. As such, retention research is particularly sensitive to the issue of "regression to the mean." In a nutshell, regression to the mean refers to the fact that extreme scores (such as those from students at the lowest end of an achievement scale) tend to move toward the mean on subsequent measures. Because students at extreme ends may subsequently score better as an artifact of measurement – rather than due to real academic growth – gains calculated on the basis of their scores cannot be assured to be valid representations of true growth. In other words, regression to the mean leads to the incomparability of gain scores. In cases in which regression to the mean may operate, analyses based on gain scores cannot be taken to yield estimates of factors relating to true growth in academic achievement.

The authors make no note of regression to the mean; however, a cursory look at their statistical results reveals that it is very much a factor. Consider the authors' first regression analysis, the results of which are included in their Table 3, and presented here below.[12]

Table 3: Effect of Being Subject to Retention Policy on FCAT Reading Test			
	Effect on FCAT Reading Test	Standard Error	P-Value
Student Is Subject to Policy	16.66	1.92	0.0000
American Indian	1.91	20.09	0.9242
Asian or Pacific Islander	36.56	9.67	0.0002
Black, Not Hispanic	-38.67	2.50	0.0000
Hispanic	-3.48	3.20	0.2768
Multiracial	14.95	7.31	0.0409
Receives Either Free or Reduced-Price Lunch	-54.81	2.42	0.0000
Student Is Limited English Proficient	-2.33	2.87	0.4169
Baseline FCAT Reading Test Score	-0.48	0.00	0.0000
Adjusted R-Squared	0.161		
N	89604		

The authors report the positive coefficient for "Student is Subject to Policy" and interpret it as indicative of a positive effect for retention. What is striking, however, is that this same model predicts a strong negative relationship between baseline FCAT Reading test scores and growth in FCAT reading test scores. According to this model, regardless of ethnicity, free/reduced-price lunch eligibility, English proficiency, or whether the policy was in place, students scoring low (relative to their peers) one year were expected to make much greater gains over the subsequent year.

The table inexplicably omits the intercept estimated for this statistical model, making it impossible to compare the projected growth of different students. In response to a personal request, the authors graciously provided these intercept estimates; the intercept reported for the table above was 730.2.[13] Using this intercept and the results reported in the table, we can shed some light on the statistical model's prediction of FCAT growth for different students. Consider two students from the 2003 cohort – one who scored 1100 in FCAT reading in the baseline year and another who scored 200. The authors' statistical model suggests that the higher-scoring student would gain 218 points over the year, compared to a whopping 651 point gain for the lower-scoring student. Furthermore, this 433-point difference in gains would be expected *regardless* of whether a student was subject to retention or not. The statistical model predicts that lower-performing students substantially outgain higher-performing students.

If this were actually happening in Florida, then scores between higher- and lower-scoring students would be converging; the state would provide a startling national model for closing the so-called 'achievement gap.' But the unfortunate truth is that Florida's achievement gap trends are not substantially better than those in other states; the model is simply not reflective of the reality of achievement trends in Florida. The differential growth reflected in the authors' statistical analyses has more to do with the nature of the score scale – that is, the tremendous range between low and high scores within the lowest performance level – than it does any true pattern in student achievement. Regression to the mean is clearly a major concern in this analysis, but the authors make no mention of it. In fact, the same relationship between baseline scores and subsequent gains is apparent in every single analysis reported in the study – for reading as well as math scores, for Stanford 9 as well as FCAT scores.

4. *The authors fail to include a key variable in their statistical analysis: the in-*

teraction between retention and initial level of achievement.

The authors fall victim to a mistake that has long led to inappropriate conclusions from regression-based analyses: they interpret a single regression coefficient yet disregard the rest of the statistical model that generated that coefficient. Two of the 20th century's most respected statisticians – Princeton's John Tukey and Harvard's Fred Mosteller – warned statisticians against this mistake in their classic chapter on "Woes of Regression Coefficients":[14]

> …a coefficient in a multiple regression – either in a theory or in a fit – depends on MORE than just:
> - the set of data and the method of fitting.
> - the [variable] it multiples.
>
> It also depends on:
> - what else is offered as part of the fit.

Simply put, in their enthusiasm concerning the positive effect they found for retention, the authors failed to recognize that the validity of this effect was conditional on (a) the validity of the other effects estimated by that statistical analysis, and (b) the nonexistence of additional variables that should have been included (which may have changed the interpretation of any or all of the current predictors).

It has already been noted above that the interpretation of the small effect for retention requires that regression to the mean also be acknowledged (because of the substantial negative relationship between baseline scores and gain scores). A second problem is the omission of a key variable of interest – the interaction between the retention variable and the baseline score. Regression courses typically instruct students to investigate interactions between variables to verify whether a given effect (such as the effect of retention) is the same across all levels of other variables. In the case of the Greene & Winters report, including the interaction between the baseline score variable and the retention variable would do just that. Furthermore, doing so would allow an additional, very significant policy question to be addressed: Is the effect of the retention policy the same for all students, or are students at different levels of ability affected differently? Clearly is important thing to know – not only will this question provide more nuanced understanding of the effects of retention, but it could also change the magnitude of the main effect estimated for the retention variable (on which most of the authors conclusions are based). Yet this important effect is excluded from the authors' analyses. It should be noted that interactions would also facilitate the exploration of other important policy questions, such as whether the effect associated with retention policy is the same or different for students who vary in terms of ethnicity, socioeconomic status, and language proficiency. Again, these important questions are left unaddressed by the authors' analyses.

In sum, to believe the validity of any single regression coefficient (such as the estimated effect of retention) one must believe in the accuracy of the statistical model that produced it; that a major effect is excluded from their analyses (even though the authors needed no additional information to include it) calls into question any inferences the authors make on the basis of those analyses.

5. *The authors report only a fraction of the evidence necessary to determine the accuracy of the statistical results.*

Many instances of insufficient reporting of analytical results have been noted above. In general, responsible reporting of results of

regression analyses should include the following:

- Descriptive statistics and/or graphs detailing the distribution of each independent and dependent variable in terms of central tendency, variability, distribution shape, and departures from this shape (i.e., outliers).
- The results of several alternative models testing the predictive strength of alternative collections of independent variables (not just a single regression analysis).
- Estimates, standard errors, and significance values for each coefficient included in each model.
- Omnibus statistics such as F-tests (and F-tests of alternative models) and R^2 values.
- Verification that the assumptions underlying each model (or at least the model ultimately selected) have been met.
- Interpretation of the collection of estimates generated by the model ultimately selected (rather than just a single estimate).

The current study falls short on providing the above results. No descriptive statistics are provided. Only a single model is presented; this model excludes intercept estimates as well as any interactions. R^2 values are provided; F-test results are not. Assumptions are not mentioned, and only a single effect (the retention effect) is interpreted (independent of other important effects such as the negative coefficient of baseline scores which suggests regression to the mean).

VI. REVIEW OF THE VALIDITY OF FINDINGS AND CONCLUSIONS

The conclusions drawn in these papers are not supported by the statistical evidence provided in those papers. All of the authors'

conclusions regarding the impact of retention on achievement are based on their statistical models. To believe the validity of any single regression coefficient (such as the estimated effect of retention) one must believe in the accuracy of the statistical model that produced it. Evidence supporting the accuracy of their statistical models is so scant, and the influence of threats to validity is so obvious, that none of the authors' conclusions can be considered valid in the absence of substantial additional information. This is not to say that future analyses might not demonstrate positive effects of retention policies; rather, the analyses presented by the authors provide no evidence to support such a conclusion.

VII. THE REPORTS' USEFULNESS FOR GUIDANCE OF POLICY AND PRACTICE

The conclusions reported in the authors' two related papers are without warrant and should not be relied upon to guide policy or practice. Policy makers reading the papers should exercise caution in two areas: relative to the current study and relative to retention studies in general. The first category is straightforward – the authors' results are based on questionable statistical analyses rife with threats to validity and therefore any claims made on the basis of these analyses are without warrant. Many of the questions raised in this review could have been answered given simple descriptive statistics on the baseline tests; the authors omit this crucial evidence from their report. At the very least, policy makers should demand that the authors release this evidence before they give any weight to the authors' conclusions.

The second level of advice reaches beyond the current study. The challenges of retention research are well established. Such research is typically retrospective; randomization of students into control and treatment

groups – the most accepted design for supporting causal inferences – is unlikely and ethically not feasible. Regression to the mean is a statistical artifact most likely in cases involving extreme ends of a numerical distribution – such as the lowest scorers on an achievement test (those most likely to be subject to retention policies). This artifact should always be examined and taken into account when appropriate. Test score gains at different places on the score scale may not be comparable instructionally.[15]

All of the above are examples of problems that could threaten the validity of retention studies. The responsible policy maker is advised to become familiar with many of these challenges, as each may threaten a given study's ability to identify the positive or negative effects of retention.

NOTES & REFERENCES

[1] Unless otherwise noted, page references are to the Manhattan Institute report.

[2] Greene, J. & Winters, M. (2004, December). *An Evaluation of Florida's Program to End Social Promotion.* Education Working Paper. New York, NY: Manhattan Institute, p. 6.

[3] By comparison, in 2003 just over 188,000 third grade students completed the FCAT.

[4] Holmes, C. T. (1989). "Grade Level Retention Effects: A Meta-Analysis of Research Studies," in Eds. Lorrie A. Shepard and Mary Lee Smith, *Flunking Grades: Research and Policies on Retention*. The Falmer Press.

[5] Nagoaka, J. & Roderick, M. (2004). *Ending Social Promotion: The Effects of Retention.* Consortium on Chicago School Research.

[6] Shepard, L.A., Smith, M.L., & Marion, S.F. (1996). Failed Evidence on Grade Retention. *Psychology in the Schools 33*(3), pp. 251-261.

[7] Greene, J. & Winters, M. (2004, December). *An Evaluation of Florida's Program to End Social Promotion.* Education Working Paper. New York, NY: Manhattan Institute, p. 6.

[8] Marcus Winters (personal communication, February 8, 2006).

[9] Florida Department of Education. (2004). *FCAT Reading Scores Statewide Comparison for 2001 to 2005.* Retrieved February 4, 2006, from http://fcat.fldoe.org/mediapacket/pdf/05gr410_Statewide_Comparison_Reading.pdf

[10] Florida Department of Education. (2004). *Parent Information – Student Report.* Retrieved February 4, 2006, from http://fcat.fldoe.org/2004/pdf/parentinfo.pdf

[11] Florida Department of Education. (2004). *Understanding FCAT Reports 2004.* Retrieved February 4, 2006, from http://www.firn.edu/doe/sas/fcat/pdf/fc_ufr2004.pdf

[12] Greene, J. & Winters, M. (2004, December). *An Evaluation of Florida's Program to End Social Promotion.* Education Working Paper. New York, NY: Manhattan Institute, p. 12.

[13] Marcus Winters (personal communication, February 8, 2006).

[14] Mosteller, F. & Tukey, J.W. (1977). *Data analysis and regression – a second course in statistics.* Philipines: Addison-Wesley, p. 300.

[15] These and other issues are discussed in

Shepard, L.A., Smith, M.L., & Marion, S.F. (1996). Failed Evidence on Grade Retention. *Psychology in the Schools 33*(3), pp. 251-261.

Shepard et al. also cite several seminal studies on grade retention.

This document is available on the Education Policy Studies Laboratory website at:
http://www.asu.edu/educ/epsl/EPRU/ttreviews/EPSL-0602-119-EPRU.pdf