Abstract Title Page

Accountability Pressure, Academic Standards, and Educational Triage

Douglas Lee Lauen, Ph.D MPP
Assistant Professor of Public Policy
University of North Carolina at Chapel Hill
dlauen@unc.edu

S. Michael Gaddis
University of North Carolina at Chapel Hill
mgaddis@email.unc.edu

Accountability Pressure, Academic Standards, and Educational Triage

Abstract Body

Background / Context: Educational accountability is in flux. Due to the failure of Congress to reauthorize NCLB, the Obama administration will soon grant waivers from the Act's requirements. Therefore, it appears that states will gain some flexibility over how to hold schools and teachers accountable, provided they establish meaningful accountability while also raising standards. This mix of accountability and increasing standards could have unintended consequences for a phenomenon called educational triage. Status-based accountability systems such as NCLB hold schools accountable for test score levels (rather than growth or gain). In these systems, schools may face incentives to adopt triage practices, which involve assessing students' likelihood to pass or fail a test and then diverting resources to those within reach of a passing test score, thus giving rise to what has become known as "educational triage" (e.g. Booher-Jennings, 2005; Brown & Clift, 2010). This could produce larger accountability treatment effects for students near grade level than students well below or well above grade level. The aim of this study is to determine whether educational triage becomes more pronounced when proficiency standards increase.

Prior research shows that accountability pressure tends to increase student test scores (Carnoy & Loeb, 2002; Chiang, 2009; Dee & Jacob, 2011; Figlio & Loeb, 2011; Figlio & Rouse, 2006; Hanushek & Raymond, 2005; Jacob, 2005; Jacob & Lefgren, 2004; Reback, 2008; Reback, Rockoff, & Schwartz, 2011; Wong, Cook, & Steiner, 2009). A growing, but contradictory, literature examines the heterogeneity of treatment effects from accountability pressure, focusing particularly on whether all students will get equal benefits from accountability pressure (Booher-Jennings, 2005; Ladd & Lauen, 2010; Neal & Schanzenbach, 2010). In a qualitative study from Texas, Booher-Jennings (2005) found that the ideology of data-driven decision making led teachers to focus instruction on the bubble kids, that is, those students just below grade level proficiency. Triage theory predicts larger accountability-induced test score gains for students near test score proficiency cut scores for students well below or well above the cut score. Quantitative studies of NCLB accountability from Washington State (Krieg, 2008) and of district-level and NCLB accountability from Chicago (Neal & Schanzenbach, 2010) report larger increases for students in the middle of the test score distribution than for low or high achieving students. Other quantitative studies find no evidence of educational triage favoring students in the middle of the test score distribution (Dee & Jacob, 2011; Ladd & Lauen, 2010; Reback, 2008; Reback, et al., 2011; Springer, 2008).

Overlooked in existing research on educational triage is whether increasing educational standards and proficiency cut scores could increase incentives to triage. Consider two extreme scenarios. State A sets the proficiency level at the 5th percentile. State B sets the proficiency level at the 65th percentile. Because there are so few students below grade level in state A, we might not expect accountability pressure to lead to large disparities in test score increases between students near grade level and students below grade level. Because standards are set so low in state A, however, we might expect a gap in the accountability-induced disparities between high and average achieving students. With many more students below grade level, the potential for accountability pressure to lead to a gap between middle and low achieving students is much greater in state B than in state A, whereas the potential for an accountability-induced gap between average and high achieving students is lower.

Purpose / Objective / Research Question / Focus of Study: this study aims to determine whether educational accountability promotes educational triage. This study exploits a natural experiment in North Carolina in which standards increased first in math in 2006 and second in reading in 2008 to determine whether an increase in educational standards caused an increase in educational triage at the expense of low and high achievers and to the benefit of students near grade level. We hypothesize that changing the rigor of state academic standards will have heterogeneous effects on student achievement. Triage theory predicts that under a status based system such as NCLB, students near grade level in schools that fail AYP will have larger increases than students well below or well above grade level. We hypothesize that the disparities between average achievers and low achievers and average achievers and high achievers will widen when the standards for defining grade level increase. In other words, triage will become worse when the tests become more difficult and the proficiency cut score increases.

Setting: We conduct our study on the universe of North Carolina public schools. North Carolina had relatively low academic standards until the mid to late 2000s. In response to the wide disparity between state and NAEP test score results, policymakers increased standards first in math in 2006 and second in reading in 2008.
Figure **1** shows the effect raising standards had on the percent of students at or above grade level in both subjects. In the year prior to the change in standards, 89% of students were proficient in math and 88% were proficient in reading. In the year subsequent to the change in standards, percent proficient in math and reading fell to 64 and 59, respectively.

Population / Participants / Subjects: The study uses student-level micro data to examine the effect of accountability pressure on educational triage for all public school students enrolled in 3rd through 8th grade between 2004 and 2009. The unit of analysis is a student level outcome nested within a school level treatment moderated by change in state standards and student position in the prior year's test score distribution. The full analysis sample size is about 1.9 million student-year observations nested within 7,200 school-year observations (about 500,000 unique students and 1,800 unique schools). The sample is 57% white, 29% black, 7.6% Hispanic, 14% academically gifted, 13% with special education needs, 46% eligible for free or reduced priced lunch.

Intervention / Program / Practice: The No Child Left Behind Act of 2001 (NCLB) mandates school-level accountability and public reporting of Adequate Yearly Progress (AYP). If any one of the nine federally identified subgroups (such as black, Hispanic, poor, or special needs students) fails performance targets, a school fails AYP. A school failing AYP for two consecutive years is deemed "in need of improvement" (INI). High poverty INI schools which receive federal funds (Title I schools) face an escalating set of sanctions, including offering students public school choice, tutoring, and the remote possibility of takeover or reconstitution as a charter school. Non-Title I schools face only public reporting of AYP status. Table 1 shows that North Carolina had large numbers of schools and students in treatment and control categories in both the pre and post period: 1,000-1,200 schools (400,000-500,000 students) failed AYP, 600-800 schools (170,000-210,000 students) met AYP. With a difference-in-difference-in-difference (DDD) model we examine whether the effect of this treatment on student math and reading test

scores varies in two periods (lower and higher academic standards) and across three levels of student prior achievement (well below, near, and well above grade level).

Research Design: We use a non-experimental research design in which estimates are produced from a DDD panel data model. We examine how the effect of school-based accountability pressure is moderated by plausibly exogenous changes in academic standards and student position in the prior test score distribution. We include school fixed effects in our model to remove between-school confounding that could arise from differential sorting of students into schools based on unobservables and unobservable fixed characteristics of schools such as the ability of a school to attract high quality teachers.

Data Collection and Analysis: The study uses administrative records from the North Carolina Department of Public Instruction. The dependent variables are standardized math and reading end of grade test scores. North Carolina is a particularly appropriate state for this analysis because its tests are vertically equated, interval scaled, and directly linked to a statewide curriculum, have been in place since the early 1990s, and are the outcome for which schools are held accountable under NCLB. We standardize scores by grade level, subject, and year to facilitate interpretation of the magnitude of coefficients and to account for the fact that the math test was rescaled in 2006 and the reading test was rescaled in 2008.

We use a DDD model which tests whether triage favoring students in the middle of the test score distribution--and at the expense of students well below and well above grade level--becomes more pronounced after grade level proficiency standards increase:

$$\begin{aligned} A_{ist} = {} & \beta_0 + \beta_1 T_{s,t-1} + \beta_2 Post_t + \beta_3 T_{s,t-1} XPost_t + \beta_4 M1_{i,t-1} + \beta_5 M2_{i,t-1} + \\ & \beta_6 M1_{i,t-1} 1XT_{s,t-1} + \beta_7 M2_{i,t-1} XT_{s,t-1} + \beta_8 M1_{i,t-1} XPost_t + \\ & \beta_9 M2_{i,t-1} XPost_t + \beta_{10} M1_{i,t-1} XT_{st-1} XPost_t + \beta_{11} M2_{i,t-1} XT_{st-1} XPost_t + \\ & \boldsymbol{\lambda\phi}_s + \boldsymbol{\beta X}_{ist} + e_{ist} \end{aligned} \quad (1)$$

where *A*, a continuous standardized achievement score[1] for student *i* in school *s* at time *t* is regressed on *T*, a school-level, time-varying, accountability threat 1-0 indicator variable, coded 1 if student *i*'s school failed the AYP in the prior year; *Post*, a 1-0 indicator variable coded 1 if the student test score is from a year after an increase in academic standards for a particular subject; *M1*, a 1-0 indicator variable coded 1 if student *i* is at least 0.5 SD below grade level in the prior year; *M2*, a 1-0 indicator variable coded 1 if student i is at least 0.5 SD above grade level; $\boldsymbol{\phi}_s$, a vector of school fixed effects; and $\boldsymbol{X}$[2] a vector of student characteristics. Because North Carolina went from relatively low to higher academic standards, we hypothesize that accountability threats would increase test scores for students near grade level more in the post period than in the pre-period. Therefore, β_3 should be positive. Furthermore, we expect that the increase in academic standards would increase the accountability-induced gaps between students near grade level and those well below and well above grade level. Therefore, we hypothesize that $\beta_{10} < 0$ and $\beta_{11} < 0$.

[1] We examine test score levels rather than gains because NCLB holds schools accountable for test score levels and not gains.

[2] Student-level variables included in the $\boldsymbol{X}$ vector include race, gender, gifted, special education, and LEP designations, and structural (i.e., switching to a school due to a change in grade configuration) and non-structural (i.e., switching to a school due to a residential move) school moves.

Estimating equation 1 without school fixed effects will produce consistent parameter estimates under two conditions: 1) no time-varying confounders of the outcome, and 2) no time-invariant between-school confounding. The difference-in-difference estimation approach reduces, but does not eliminate, between-school confounding. We include school fixed effects to remove between-school heterogeneity that would otherwise bias the accountability effect. This approach estimates the effect of accountability pressure on successive cohorts of students in the same school. The specification shown in equation 1 uses only within-school variation across cohorts to produce estimates of accountability pressure. Therefore, those schools consistently failing or consistently attaining performance targets contribute nothing to the accountability pressure estimates. While this approach ignores the performance of schools with static treatment statuses, we view comparing the performance of cohorts exposed to accountability pressure to cohorts not exposed to such pressure who attended the same school at different points in time as more valid than comparing the performance of students in different schools. Time-varying school-level confounders that differ in the pre and post periods remain threats to the validity of our findings. Including school fixed effects buys us some, but not complete, protection from all possible sources confounding.

Findings / Results: Figure 2 displays the primary parameters of interest from the school fixed effects DDD regressions shown in columns 2 and 4 of Table 2. Consistent with the triage hypothesis, the figure shows positive effects for students at the margin of grade level (.216 SD in math and .068 SD in reading) and negative differential effects for students well below and well above grade level, with disparities larger in math than in reading. In math there is strong evidence that the increase in academic standards benefitted students near grade level more than low or high achieving students. In reading, the evidence is somewhat weaker, but it is clear that high achieving students benefitted less than students near or below grade level. In the complete version of this paper we plan to examine whether these effects vary by the number below-grade-level students a school serves. In addition, we plan to present predicted values for low, average, and high achievers in both the pre and post period to more clearly assess the differential effects of increasing standards.

Conclusions: As states move to increase academic rigor by adopting the Common Core Standards while also implementing changes in accountability policy, it is critical for education research on educational accountability from the NCLB era to inform policy design. Research has clearly shown small, but measurable, benefits from school-based accountability pressure. Research on the heterogeneity of school-based accountability pressure on students at different points in the prior achievement distribution is mixed, but has overlooked the role the rigor of academic standards may play in promoting or reducing the potential for educational triage. Our view is that school-based accountability based on test score level (instead of growth) is a necessary, but not sufficient cause of triage. In low standards states, the risk of triage harming low achieving students may be particularly low. As we have shown, however, raising academic standards can promote triage, especially in math. These findings suggest that states may want to take a gradual approach to increasing academic standards, which may be particularly important as the unit of accountability moves from the school to the individual teacher.

Appendices

Appendix A. References

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal, 42*(2), 231-268.

Brown, A. B., & Clift, J. W. (2010). The Unequal Effect of Adequate Yearly Progress: Evidence From School Visits. *American Educational Research Journal, 47*(4), 774-798.

Carnoy, M., & Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9-10), 1045-1057. doi: 10.1016/j.jpubeco.2009.06.002

Dee, T. S., & Jacob, B. (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management, Forthcoming*.

Figlio, D. N., & Loeb, S. (2011). School Accountability. In E. A. Hanushek, S. Machin & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, pp. 383-421). North-Holland, The Netherlands: Elsevier.

Figlio, D. N., & Rouse, C. E. (2006). Do Accountability and Voucher Threats Improve Low-Performing Schools? *Journal of Public Economics, 90*((1-2)), 239-255.

Hanushek, E. A., & Raymond, M. E. (2005). Does School Accountability Lead to Improved Student Performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5-6), 761-796. doi: 10.1016/j.jpubeco.2004.08.004

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement - Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources, 39*(1), 50-79.

Krieg, J. M. (2008). Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act. *Education Finance and Policy, 3*(2), 250-281.

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management, 29*(3), 426-450.

Neal, D., & Schanzenbach, D. W. (2010). Left Behind by Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics, 92*(2), 263-283.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics, 92*(5-6), 1394-1415. doi: 10.1016/j.jpubeco.2007.05.003

Reback, R., Rockoff, J., & Schwartz, H. L. (2011). *Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under NCLB.* Working Paper.

Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review, 27*(5), 556-563. doi: 10.1016/j.econedurev.2007.06.004

Wong, M., Cook, T. D., & Steiner, P. M. (2009). *No Child Left Behind: An Interim Evaluation of its Effects on Learning using two Interrupted Time Series each with its own Non-Equivalent Comparison Series.* Working paper.

Appendix B. Tables and Figures

Not included in page count.

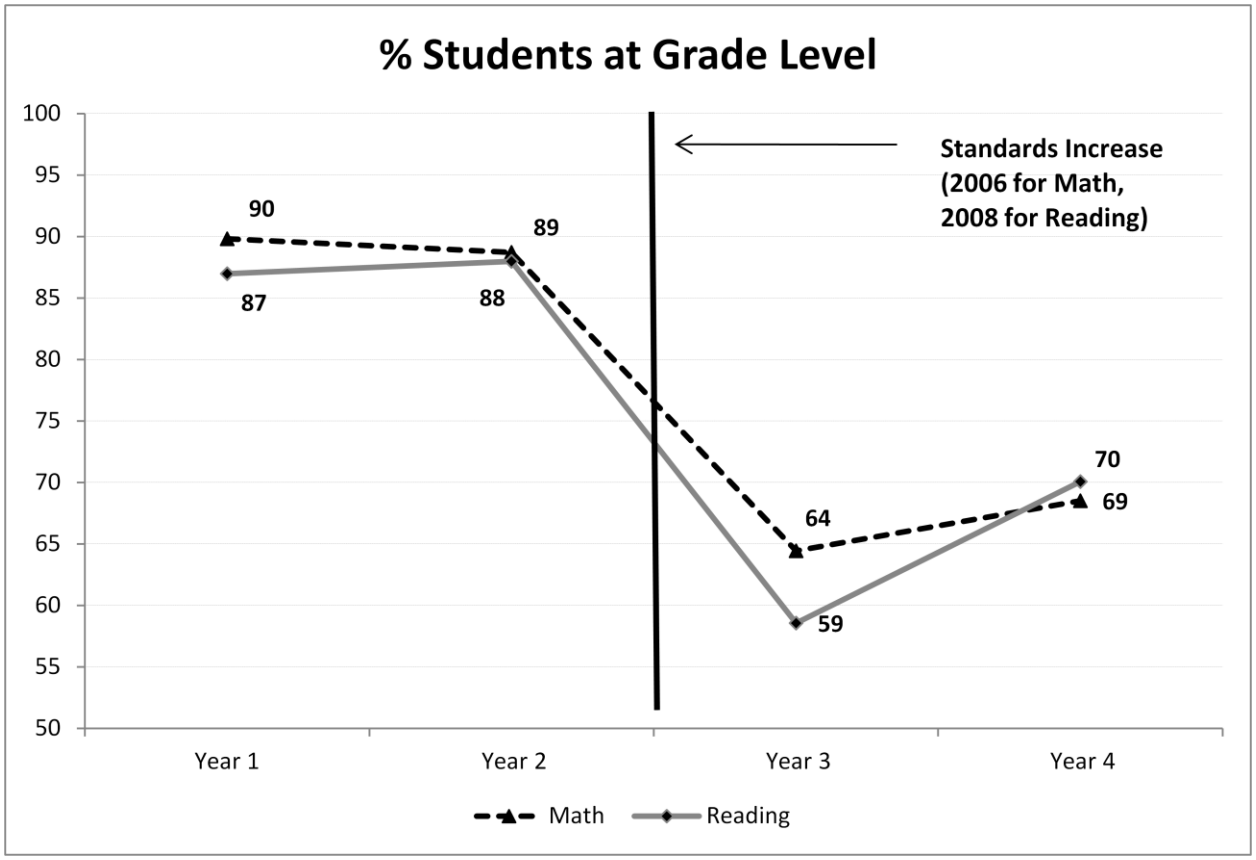

Figure 1. Percent of Students at Grade Level Pre and Post Change in Math and Reading Standards. Note: in reading year 1-year 4 represents 2006-2009; in math year 1-year 4 represent 2004-2007.

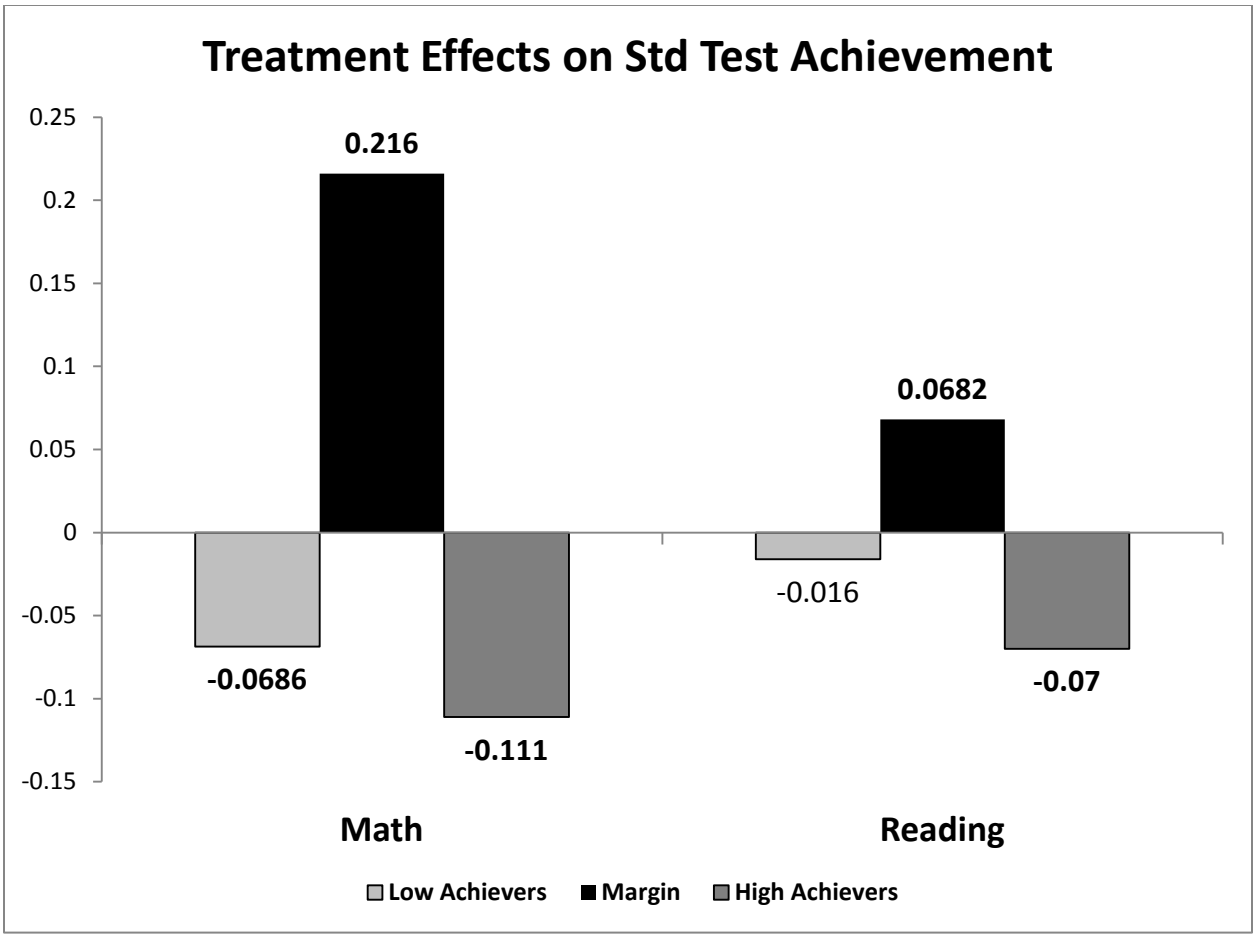

Figure 2. Treatment Effects on Standardized Test Achievement (Math and Reading). Note: Figure created using school fixed effect models (2 and 4) from table 1. Bolded coefficients are statistically significant at $p<.05$.

Table 1. Number of schools and students in schools failing AYP, by year (Reading)

	Pre		Post	
	2006	2007	2008	2009
Schools				
Met	721	749	794	636
Failed	1,054	1,043	1,008	1,181
Students				
Met	197,416	192,449	211,798	168,056
Failed	419,240	451,944	440,269	486,664

Table 2. DDD Models Predicting Standardized Achievement

	(1) Math Pooled OLS 2004-2007	(2) Math School FE 2004-2007	(3) Reading Pooled OLS 2006-2009	(4) Reading School FE 2006-2009
Black	-0.267***	-0.242***	-0.257***	-0.249***
	(0.00491)	(0.00372)	(0.00414)	(0.00346)
Hispanic	-0.0194**	-0.0145**	-0.0695***	-0.0708***
	(0.00614)	(0.00455)	(0.00482)	(0.00402)
Other Race	-0.00275	0.0194***	-0.0513***	-0.0414***
	(0.0119)	(0.00565)	(0.00941)	(0.00368)
Male	0.0386***	0.0404***	-0.0514***	-0.0509***
	(0.00152)	(0.00146)	(0.00124)	(0.00122)
Academically Gifted	0.925***	0.913***	0.749***	0.740***
	(0.00576)	(0.00443)	(0.00462)	(0.00378)
Special Education	-0.295***	-0.312***	-0.288***	-0.306***
	(0.00408)	(0.00327)	(0.00372)	(0.00334)
Limited English Proficiency	-0.171***	-0.180***	-0.206***	-0.217***
	(0.00541)	(0.00447)	(0.00446)	(0.00386)
Ever Retained	-0.263***	-0.258***	-0.231***	-0.230***
	(0.00321)	(0.00294)	(0.00347)	(0.00327)
Student Poverty	-0.210***	-0.155***	-0.203***	-0.160***
	(0.00505)	(0.00212)	(0.00339)	(0.00199)
Structural School Move	-0.0423***	-0.0721***	-0.0658***	-0.0624***
	(0.00634)	(0.00481)	(0.00442)	(0.00337)
Non-Structural School Move	-0.0569***	-0.0600***	-0.0545***	-0.0559***
	(0.00305)	(0.00245)	(0.00265)	(0.00237)
Post-Revised Standards	0.192***	0.0832***	0.369***	0.329***
	(0.0136)	(0.0105)	(0.0101)	(0.0104)
Failed AYP	0.00521	0.0666***	-0.0468***	-0.0551***
	(0.00754)	(0.00834)	(0.00691)	(0.00775)
Low Achievement	-0.358***	-0.347***	-0.488***	-0.475***
	(0.00984)	(0.00993)	(0.00957)	(0.00950)
High Achievement	0.920***	0.863***	0.899***	0.858***
	(0.00866)	(0.00665)	(0.00752)	(0.00613)
Low Ach * Post	-0.0373*	-0.0515***	-0.0741***	-0.0710***
	(0.0148)	(0.0148)	(0.0120)	(0.0116)
High Ach * Post	-0.115***	-0.0787***	-0.254***	-0.234***
	(0.0103)	(0.00920)	(0.00836)	(0.00826)
Low Ach * Failed AYP	0.0120	0.00652	-0.0472***	-0.0480***
	(0.0107)	(0.0109)	(0.0106)	(0.0105)
High Ach * Failed AYP	-0.0544***	-0.0180*	-0.0362***	-0.0146*
	(0.0114)	(0.00876)	(0.00865)	(0.00701)
Post * Failed AYP	0.119***	0.216***	0.0432***	0.0682***
	(0.0152)	(0.0118)	(0.0121)	(0.0121)
Low Ach * Post * Failed AYP	-0.0865***	-0.0686***[a]	-0.0137[b]	-0.0160[c]
	(0.0159)	(0.0159)	(0.0137)	(0.0133)
High Ach * Post * Failed AYP	-0.0768***	-0.111***[a]	-0.0550***[b]	-0.0700***[c]
	(0.0122)	(0.0106)	(0.00965)	(0.00936)
Constant	-0.589***	-0.622***	-0.433***	-0.503***
	(0.00791)	(0.00688)	(0.00666)	(0.00646)
Observations	1,917,437	1,917,437	1,906,801	1,906,801
R^2	0.560	0.524	0.540	0.496

Note: Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; a indicates two coefficients are significantly different at p<0.05; b indicates two coefficients are significantly different at p<0.01; c indicates two coefficients are significantly different at p<0.001.