Abstract Title Page
Not included in page count.

Title: Games Schools Play: How schools near the proficiency threshold respond to accountability pressures under No Child Left Behind

Author(s): Vivian C. Wong

Abstract Body

Background / Context: Although education reform has remained at the forefront of the domestic policy agenda in the United States, there is little consensus on what the policy should look like, what its aims should be, and how it should be implemented (i.e. at the state or federal level). Early reform efforts focused on compensatory education programs such as Title I that provided schools with increased resources for instruction, especially for disadvantaged and minority youth. But student achievement scores remained disappointingly low, so recent reform initiatives have focused on holding schools directly "accountable" for student outcomes. The rationale here is that by aligning achievement standards with testing and accountability measures, schools would have strong incentives to improve student performance. Most accountability measures involve annual reporting of aggregate student achievement at both the school- and district-levels, and some form of remediation for schools and districts that miss targets in proficiency subjects. Corrective actions may include increased oversight by state departments of education, additional resources for teacher professional development and student tutoring services, and possible reconstitution or closure of schools that persistently fail to meet academic performance standards.

In 2001, the Bush Administration launched a federal version of accountability reform under the auspices of No Child Left Behind (NCLB). The initiative mandated that schools receiving Title I funding were subject to remediation if they failed to meet academic proficiency targets. The goal was for all students to be considered "proficient" – however broadly defined – by 2014. To achieve this mandate, states were required to conduct annual testing on exams that were aligned with curriculum standards; establish a time schedule for which all students would become proficient by states' own proficiency standards; and impose remediation on schools in need of improvement. The program was infused with substantial funding for enforcement at both the state and federal levels. Thus far, three efforts have evaluated the effectiveness of No Child Left Behind; all have employed an interrupted time series design examining abrupt changes in NAEP scores prior to and post implementation of NCLB in 2002. Hanushek and Raymond (2005), Dee and Jacob (2009), and M. Wong, Steiner, and Cook (2009) all found positive effects of NCLB, though the latter two found only significant effects for math but not for reading.

The positive findings, however, are tempered by persistent questions on how states, schools and teachers actually respond to accountability pressures, and whether these effects are the result of concerted efforts to improve students' generalized knowledge and skills, or of some other "gaming" practice intended to raise schools' proficiency scores but not students' achievement levels. In recent years, researchers have examined many of the negative or perverse effects of accountability policies, such as practices that prevent low achieving students from entering the testing pool, or only teaching items known to be covered on high stakes tests (Diamond, 2007; Cullen & Reback; 2006; Figlio, 2006; Figlio & Getzler, 2006; Heilig & Darling-Hammond, 2008; Jacob, 2005).

Purpose / Objective / Research Question / Focus of Study: The purpose of this paper is to explore state and school responses to accountability pressures under NCLB. Section 2 presents a framework for summarizing the many positive and negative ways in which states, schools, and teachers respond to accountability pressures under NCLB. Section 3 focuses on state accountability policies that in recent years have become more relaxed in allowing schools to lower proficiency thresholds, but also more stringent in the specificity of their requirements. The

paper argues that schools with proficiency scores just below state accountability cutoffs have the most incentive to engage in gaming practices, and that these schools have capitalized on recent changes in state policies to introduce new methods for gaming. Using data from Pennsylvania, the paper provides empirical evidence that schools just below the state threshold engage in gaming practices and suggest mechanisms that these schools employ.

Intervention / Program / Practice: NCLB requires states to use five indicators to determine adequate yearly progress (AYP): 1) the percentage of students who are proficient in reading as measured by the state reading assessment; 2) the percentage of students who are proficient in mathematics as measured by the state mathematics assessment; 3) the percentage of students who participate in state reading assessments; 4) the percentage of students who participate in state mathematics assessments; and 5) at least one other academic indicator at each school level (elementary, middle, and high school). All schools held accountable must meet AYP targets for the school as a whole, and for any student subgroup that exceeds a state-set minimum number of students.

Schools that fail to meet proficiency targets in subject areas may apply exemption rules to achieve state AMOs. Although states have discretion in determining which rules to apply, two of most common are the confidence interval and safe harbor rules. Confidence intervals apply a "plus or minus" band around the state's minimum proficiency score, or the annual measurable objective (AMO). Under this rule, given the number of students in a group, a confidence interval is constructed around the AMO target, which effectively extends the cutoff to several percentage points above and below the state threshold. Thus, to make AYP, a school needs only to achieve a target score that is equivalent to the lower bound of the confidence interval. For safe harbor, a state examines a school's prior year performance and uses its "safe harbor" rule (i.e., an annual improvement of 10 percentage points of students who are not proficient in the previous year) to calculate the effective cutoff for the school or subgroup. These rules are systematically and uniformly applied to all schools in the state and are public knowledge, so all schools' site-specific cutoffs are completely observable and may be calculated by applying the states' exemption rules. However, under the confidence interval and safe harbor rules, schools may corrupt adjusted proficiency thresholds if they are able to a). control the number of students eligible for testing and/or b). alter the composition of students taking the test. For example, schools may reduce the threshold via the confidence interval rule if they decrease the number of students in the group, or add (probably high scoring) outliers in the group. Schools may game the safe harbor rule by altering the composition of students in the testing group the following year so that a fewer percentage of students are considered not proficient.

Schools or districts that do not make AYP for two consecutive years are identified for improvement. If schools continue to miss AYP after they are identified, they move into more serious stages of identification for improvement, which require changes in curriculum, governance, staffing, or other major reforms to the schools' operation. Thus, schools have strong incentives to avoid missing AYP and risk entering improvement status under NCLB. However, not all schools are affected equally under accountability pressures (Hanushek & Raymond, 2003). For example, one would expect schools with proficiency scores close to the threshold might be expected to alter their behavior more than schools further away from established critical thresholds. In particular, schools with scores just below the cutoff may feel compelled to engage in gaming practices that either raise their proficiency scores or lower required thresholds by a few points in order to pass.

Setting: This study examines 2006-07 AYP data for Pennsylvania schools from grades three through eight. Pennsylvania examines whether a school meets state requirements for proficiency, participation, and attendance in reading and math. If the school misses state targets for either participation or attendance, then it fails to make AYP for the year. If the school misses AYP only because of the subject proficiency requirement, then the confidence interval rule is applied and the state determines whether the school makes AYP using the adjusted threshold. If not, then the state applies the safe harbor rule first, followed by the safe harbor-confidence interval rule. If the school misses proficiency targets after all of three exemption rules are applied, then the school is designated as failing to make AYP. Schools may appeal state designations, but only in cases where schools believe a factual or statistical error has been made.

Population / Participants / Subjects: Because prior literature suggests that schools reclassify students as SWD to "game" the system, this paper looks at proficiency scores for the SWD subgroup only. Thus, the sample includes only public schools that are held accountable under federal NCLB policy and have an eligible SWD subgroup (schools with 40 SWDs or more). In total, 1035 public elementary and middle schools are included in the analysis sample, where 385 schools missed AYP in 2006-07 and 645 schools made the cutoff.

Research Design: One way to assess empirically whether schools corrupt proficiency thresholds is to borrow a method from the regression-discontinuity (RD) literature that assesses implementation threats to the design. In RD, units are assigned to treatment conditions on the basis of an assignment variable, cutoff score, and nothing else. Those with assignment scores below the cutoff receive the treatment (or comparison), and those with assignment scores above the cutoff receive the comparison (or treatment). Treatment effects are measured by the size of the discontinuity in the regression slope at the cutoff. In this study, schools' percent proficiency is like the assignment variable, the cutoff is the state's AMO target, and the intervention is failure to make adequate yearly progress.

However, RD requires that observations just to the left and right sides of the cutoff are exchangeable once the treatment condition and assignment variables are controlled for in the regression model. More formally, this is identified as the continuity assumption, which requires that potential outcomes are continuous on both sides of the cutoff. If, however, the treatment assignment process is public knowledge, and units have strong preferences to avoid treatment, then some sorting on the assignment variable may occur. This is akin to what happens in randomized experiments when participants have knowledge of treatment conditions and override the assignment mechanism to select into a desired treatment status. McCrary (2008) differentiates between RD designs with partial and complete manipulation of the assignment process. In partial manipulation, agents have some influence over their assignment score, but the process is not under their complete control. In complete manipulation, the assignment variable is entirely under the agent's control and the cutoff is publicly known. Under NCLB, complete manipulation by schools poses a serious concern. The assignment process for failure to make AYP is well known, and schools are under strong pressure to show that they are proficient. If schools are able to manipulate or misrepresent information that would affect their proficiency scores and/or site-specific cutoffs, then corruption of the proficiency scores around the threshold is likely to occur. In RD, one test for sorting is to examine whether there are discontinuities in the density function of the assignment variable. I assume that units near the cutoff have the

strongest incentives to manipulate the assignment process, so I look for any discontinuities located at the cutoff.

Data Collection and Analysis: If schools sort themselves around the cutoff, then one would expect to see a dip in the density of observations immediately below the cutoff followed by a sharp increase in the density of schools at or above the cutoff. One challenge with the Pennsylvania data is that the majority of schools made AYP via an exception rule. Because the exemption rules provide schools with their own site- and subject-specific thresholds, plotting histograms or kernel density plots of raw proficiency scores for reading and mathematics are not useful for detecting discontinuities in the density function at the cutoffs. To address this concern, the paper considers Pennsylvania's exemption rules and reading and math proficiency scores as multiple assignment rules. It then uses the "centering procedure" (as described by V.C. Wong, Steiner, & Cook, under review) to collapse multiple assignment rules into a single centered assignment variable, thereby reducing a high-dimensional assignment procedure into a single assignment mechanism.

To assess whether manipulation of the assignment score occurred around the cutoff, the paper examines kernel density estimates around the cutoff. Density estimation can be construed as an attempt to estimate the probability density function of a variable based on a sample. It can also be thought of informally as a descriptive technique for smoothing histograms. The general kernel density estimator is given by:

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

where x is the value of the assignment variable for which we want to obtain the kernel density estimate, n is the number of observations, h is the half-width of the kernel, X_i is the value of the assignment score for school i, and K is the kernel function. In this case, the triangle kernel was used. To select the optimal bandwidth, I started with a width size that produced the minimum mean integrated square error, and then tried various bandwidths by trial and error. This was to ensure that the window widths were small enough to reveal detail in the plots, but large enough to suppress random noise (for more about density estimation, see Fox, 2008).

McCrary (2008) proposed a density test that is implemented as a Wald test of the null hypothesis that the discontinuity is zero. The test involves a smoothed histogram using an extension of local linear regression. The procedure is conducted by obtaining a finely graded histogram first, and then by using local linear regression to smooth out the histogram on both sides of the cutoff.

Findings / Results: Table 1 summarizes the number of schools that made AYP via the exemption rules for reading and math. It shows that only 58 schools made AYP by meeting the state AMO target for reading and math, and that safe harbor was the most popular exemption rule for making AYP (305 schools for reading and 223 schools for math). Safe harbor-confidence interval was the next most popular rule applied for making AYP (287 schools for reading and 178 schools for math), followed by the confidence interval rule (76 schools for reading and 168 schools for math).

Figure 1 is a kernel density plot of the centered proficiency score for schools in Pennsylvania. Proficiency thresholds are based on the minimum centered value adjusted for all exemption rules for reading and mathematics. The plot shows a clear dip in the density of schools immediately before the cutoff, followed by a sharp increase just over the cutoff. The

spike in proficiency scores consists of many more schools than should be there if no sorting had occurred near the cutoff. The dip below the cutoff suggests that schools scoring right under the threshold manipulated their assignment scores in order to switch over to the comparison side.

The paper considers how schools might actually sort themselves at the proficiency thresholds. In general, schools appear to have at least two options for manipulating their assignment or cutoff scores. First, they may increase the percentage of SWDs who are "proficient" by altering the composition of students in the subgroup. Second, in states where schools can make AYP through a confidence interval rule, they may control the number of SWDs assessed or include students with high outlier scores to increase the width of the confidence interval. This would result in lower threshold scores for making AYP. Subsequent analyses indicate that schools may be manipulating the size of the SWD subgroups to lower proficiency cutoffs via the confidence interval exemption rule. However, there was no evidence that schools were gaming the safe harbor rule in Pennsylvania.

Conclusions: This study shows that in at least one state, Pennsylvania, the vast majority of schools that make AYP do so with the aid of exemption rules. Most schools pass by taking advantage of the two most generous exemption policies, the safe harbor and safe harbor-confidence interval rules. On their own, these findings do not indicate any type of illegal "gaming" occurring at the school level. Rather, it shows that the majority of Pennsylvania schools benefit from state policies that effectively reduce performance standards for proficiency under NCLB. The larger concern is evidence suggesting that schools scoring just below proficiency thresholds manipulate their assignment or cutoff scores in order to make AYP. This was indicated by the large discontinuity in the kernel density plot for schools in Pennsylvania, which showed a dip in the density of schools immediately below the cutoff followed by a spike at or above the proficiency threshold. Preliminary analysis of school AYP data for Texas shows a similar pattern in the distribution of schools, so the Pennsylvania finding is replicated in at least one other state.

Results presented in this paper diverge from earlier findings on school gaming behaviors in at least two ways. First, prior work focused on gaming activities such as retention, suspension, or drop out in generally low achieving schools. Although these activities would result in shifts in the distribution of schools, they would not explain the discontinuity in the density of schools that occurs *exactly at the cutoff*. For such manipulation to occur, at least three specific requirements must be met. First, schools must have knowledge of the cutoff. Second, they must have means to either raise their proficiency scores by a few points, or to lower the required threshold by gaming the exemption rules. Third, schools must have strong motivation to avoid or participate in the treatment condition. My investigation of Pennsylvania policies suggests two plausible scenarios in which schools game the proficiency thresholds. First, schools near the cutoff may have legally manipulated their proficiency rates by using the federal caps rules. They may have done this by optimizing the number and composition of students who were counted as proficient by using results from alternative exams. Second, schools may have illegally manipulated their site-specific cutoffs by reducing the number of SWDs assessed and/or including outliers in the subgroup, thereby increasing the width of their confidence intervals. Unfortunately, because of limitations in the school level data, this paper is only able to investigate empirically whether schools are manipulating subgroup size to alter the width of the confidence interval. Finally, the data show no evidence of schools gaming proficiency thresholds via the safe harbor rule, which serves as a strong contrast for comparing discontinuities in the density function near the cutoff.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Cullen, J. B., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. *NBER Working Paper 12286*.

Dee, T., & Jacob, B. (2009). The impact of No Child Left Behind on student achievement. *NBER Working Paper No. 15531*.

Diamond, J. B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education, 80*(October), 285-313.

Figlio, D. N. (2006). Testing, crime, and punishment. *Journal of Public Economics, 90*(4-5), 837-851.

Figlio, D. N., & Getzler, L. S. (2006). Accountability, ability, and disability: Gaming the system. In T. Gronberg & D. Jansen (Eds.), *Advances in Microeconomics*. Amsterdam: Elsevier.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. California: Sage Publications, Inc.

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.

Heilig, J. V., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis, 30*(2), 75-110.

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.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*, 698-714.

Wong, M., Cook, T. D., & Steiner, P. M. (2009). No Child Left Behind: An interim evaluation of its effects on learniing using two interrupted time series each with its own non-equivalent comparison series. *Institute for Policy Research Working Paper: WP-09-11*.

Wong, V. C., Steiner, P. M., & Cook, T. D. (under review). Analyzing Regression-Discontinuity Designs with Multiple Assignment Variables: A Comparative Study of Four Estimation Methods.

Appendix B. Tables and Figures

Not included in page count.

Figure 1: Kernel density and histogram plots of assignment variable at the cutoff for Pennsylvania

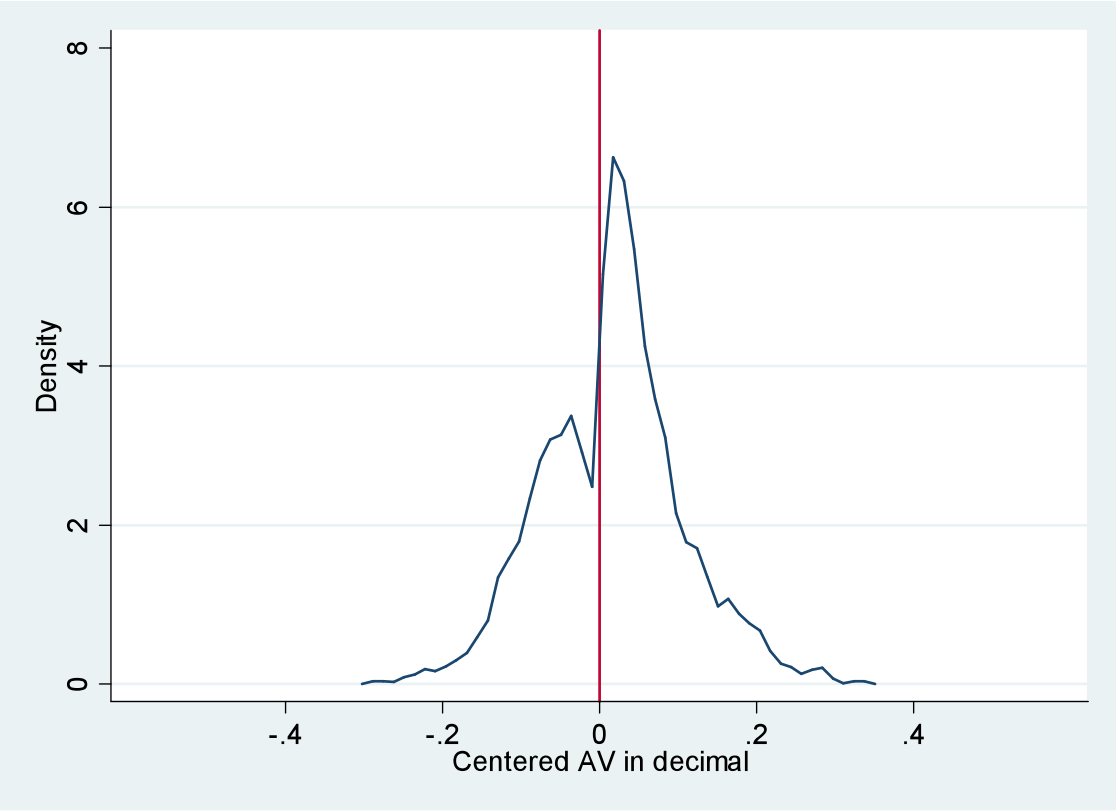

Table 1: Number of Pennsylvania schools with an SWD subgroup that made AYP via exemption rules in 2006-2007.

	Reading proficiency					
Math proficiency	Met AYP target	Met AYP by CI	Met AYP by SH	Met AYP by SH CI	Did not make AYP	Total
Met AYP target	58	55	44	36	16	209
Met AYP by CI	8	16	64	55	25	168
Met by SH	0	4	119	70	30	223
Met by SH & CI	0	1	54	61	62	178
Did not make AYP	0	0	24	65	163	252
Total	66	76	305	287	296	1,030

SH: Safe Harbor
CI: Confidence interval
SH CI: Confidence