“Going Down With the Ship?”

The Effect of School Accountability on the Distribution of Teacher Experience in California

David P. Sims
Brigham Young University
davesims@byu.edu

Working Paper

**Please do not quote or cite without permission*

Paper presented at the *NCLB: Emerging Findings Research Conference* at the Urban Institute, Washington, D.C. on August 12, 2009. The conference was supported, in part, by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), funded by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education and the National Center for Performance Incentives (NCPI) at Vanderbilt University. The author wishes to thank Chad Lee for valuable research assistance as well as the participants at the 2008 *American Education Finance Association* meetings for helpful comments and suggestions. The views expressed in the paper are solely those of the author and may not reflect those of the funders or supporting organizations. Any errors are attributable to the author.

Abstract

Many school accountability programs, including the No Child Left Behind (NCLB) A ct are built on the premise that the threat of sanctions attached to failure will produce higher student achievement. However, the stigma associated with failing schools and the expected costs of possible future sanctions may lead experienced teachers to leave these schools for other opportunities. This may undermine the program's improvement efforts. Particularly it may lead failing schools to rely on a higher proportion of novice teachers. This study looks at elementary and secondary schools in California from 2002-2006 to determine the effect of failing to meet academic performance thresholds on teacher experience under the NCLB accountability system. Because failing schools differ in important ways from schools that meet performance targets, the author takes advantage of the racial subgroup rules to compare groups of schools that may have different failure probabilities despite similar profiles. The author finds that failure to meet AYP is associated with decreases in aggregate teacher experience and increases in the proportion of novice teachers.

I. Introduction

The implementation of the federal No Child Left Behind Act of 2001 punctuated a decade of school accountability reforms. At the heart of this movement is the belief that additional incentives for school employees will lead to better outcomes from students. This policy of punishing schools for measured student failures has been controversial in education circles. Critics have charged that it promotes non-optimal use of teaching time by concentrating on activities that may improve test scores but not learning, and that it leads teachers to ignore students who are not near proficiency cutoffs (Neal and Schanzenback forthcoming; Ballou and Springer 2008). Previous academic studies have also found that high stakes incentive programs may lead to other undesirable outcomes such as teacher cheating (Jacob and Levitt 2003) and district tampering with the school calendar (Sims 2008).

Most importantly, it is unclear whether a failure based accountability program actually improves student learning. Past research has suggested that states which adopt accountability systems see higher gains on the National Assessment of Education Progress (Carnoy and Loeb 2002). Early researchers of state accountability performance found large effects ascribed to fear of stigma in North Carolina (Ladd and Glennie 2001) and due to voucher threats in Florida (Greene 2001).

In particular, one way an accountability program might undermine future student achievement is if it leads experienced teachers to transfer away from failing schools. Teachers might do this if they perceive a stigma attached to working at a failing school or if they fear that the educational interventions provided for failing schools will impose personal costs on them, such as requiring more time and effort, or increased administrator interference. Since teacher experience in the first few years is one of the few observable characteristics linked to student achievement (Goldhaber and Brewer 1997, Hanushek 1997), this could have important implications for student achievement. Indeed, recent research by Sims (2007) suggests that accountability failure due to an additional significant subgroup translates into lower student achievement on a low-stakes nationally normed test. Reduced teacher experience could partially explain this result.

However, evaluating accountability programs that punish underperforming schools remains difficult because of the lack of a meaningful control group. Schools that fail are, as a group, different in many salient ways than those that do not. In a reexamination of the Florida accountability program, Figlio and Rouse (2006) find that most of the effect of the program on achievement disappears when lagged test scores and more stringent controls are added. Because of these difficulties, there have to this point been few attempts to look at the consequences of accountability failure on teacher experience.

This paper uses the variation provided by racial subgroup rules in the California adoption of NCLB to find schools with profiles that have different accountability failure probabilities (Kane and Staiger 2002, 2003). I show that under the AYP regime the likelihood of failing to meet the acceptable standard increases sharply for California schools with extra demographic subgroups. I find that the California implementation of No Child Left Behind results in a negative relationship between failure and teacher experience, even when incorporating the variation in failure probabilities from small changes in school composition due to subgroup rules. Alternatively, from a policy perspective, this may be interpreted as a negative effect of policies that enforce subgroup accountability on teacher experience.

The remainder of the paper proceeds as follows; Section II gives some accountability background information, section III discusses the data, section IV explains the empirical strategy, section V presents the results, and section VI concludes.

II. Background

On January 8, 2002 President George W. Bush signed the No Child Left Behind Act (NCLB) which overlaid and at times superseded accountability programs in the states. Although the law in its entirety is complex, for the purposes of this paper there are a few features that are of particular note. First, though states were allowed to pick out a time dependent cutoff for the federally mandated standard of adequate yearly progress (AYP), and use their own test instruments to measure it, the standard was to be uniform for all schools (at a point in time) rather than relative to the schools immediately prior achievement.

Secondly, the NCLB regime was not timid about categorizing schools as failing. This was deliberate as social stigma was designed to motivate schools to perform better. More serious sanctions were reserved for schools that failed to make AYP in a second year. Failing schools face sanctions that range from being put on the public watch list to the possibility of staff reorganization and state takeover of the school. In practice, the actual interventions in early years most often involve professional teams giving mandatory advice about school reform and the provision of certain supplemental services to students. From the point of view of the staff, however, these sanctions involve paperwork, increased supervision and a loss of local classroom control, as well as the public labeling of your school as underperforming. Finally, the federal program imposes a school choice requirement on Title I schools that repeatedly failed to meet the standard.

While the achievement of the entire student body was necessary to pass the AYP standard, it might be insufficient for some schools. The incentive program also provided that all "numerically significant" student subgroups had to meet the schools AYP target, though it allowed states to follow self-determined numerical subgroup definitions. According to California law, the subgroup designation applied to a racial minority or socio-economically disadvantaged group of students, "that constitutes at least fifteen percent of a school's total pupil population and consists of at least thirty pupils" (50 after the first year). Furthermore any subgroup with at least 100 pupils is numerically significant no matter what percentage of the schools population they represent (PSAA 1999). Similar provisions exist in the accountability laws of other states. The intent of this provision was likely an attempt to force schools to pay attention to the needs of all their students. However it also creates a double standard whereby schools with slightly different student bodies and similar test scores face widely different probabilities of failing to meet the standard.

In theory, an accountability regime that punishes repeat failures can influence schools in two ways. First, there is an incentive to avoid any stigma and not fail in the first place. Second, there is a further incentive for teachers at failing schools to leave, thereby avoiding serious sanctions and further public scorn. Since most California districts heavily weight experience in their transfer policies, the experienced teachers would likely have more ability to move, should they wish to do so. Unfortunately, my approach in this paper is only capable of capturing behavior through the second channel.

III. Data

The data for this project comes from the California Department of Education. The CDE maintains files on each school's AYP status in all school years from 2001-2002 onwards. Additionally, these files give school level breakdowns of race, socioeconomic status, free lunch, parents education, English learners and so forth. Thus for each year I can construct control variables for school characteristics and match them with a school's accountability performance as well as information about student subgroups. I have obtained data from these files for California schools for the three school years beginning in the fall of 2002.

The Educational Demographics office also maintains teacher level information gathered from a yearly (fall) data collection involving all public school teachers in the state. This Professional Assignment Information Form includes information about each teacher's education, experience and credentialing status. For the purposes of this study I have collapsed this data to school-year averages to match it with the accountability data.

The California data have a small sample trimming issue since test score data is only reported for schools with at least 100 students, causing much of my data to completely miss small schools. I have further excluded from the sample schools for which my desired student characteristic covariates are unavailable, though this does not seem to have a meaningful effect on my results.

Descriptive statistics for the data are given in Table 1. Interestingly, California classrooms have a relatively large proportion, 16%, of novice teachers, here defined as those in the first three years of teaching. The average teacher has about thirteen years of experience in teaching with just over ten of those years coming with their current district. This discrepancy supports the notion that transfers between districts are common but unlikely ubiquitous. California also has a high degree of racial and socioeconomic diversity. The average school in this time period had half of its students eligible for federal meal subsidies and had a forty-three percent Hispanic student population. Furthermore the average school has two and a half significant subgroups (in addition to its majority group), and over ten percent of the schools have at least four extra subgroups. The most common subgroup is Hispanic students, which occurs in more than 70 percent of schools. Given the extra subgroups it is, perhaps, unsurprising that the school failure rate is above the NCLB first and second year national average of 26 percent.

In order to obtain consistent estimates of the effect of accountability failure on teacher experience, I will use subsamples of this data that reflect groups of schools that are relatively more comparable than the schools in California as a whole. Descriptive statistics for the two most important samples of this data are also shown in Table 1. In general, these samples diverge somewhat from the state as a whole. The schools that have between ten and twenty percent Hispanic students have higher teacher experience and socioeconomic status indicators while those with between ten and twenty percent black students have lower experience and higher failure rates.

IV. Empirical Strategy

Given my desire to estimate the effect of school failure on subsequent teacher experience, I adopt a statistical model of the following form:

$$E_{it} = \varphi F_{it-1} + X'_{it}\gamma + \mu_{it} \qquad (1)$$

where i indexes schools and t time. F_{it-1} is an indicator for failure to meet the standard in the previous year, the achievement variable E_{it} is a measure of school level teacher experience, and X is a vector of school average controls to account for student characteristics including race, free lunch and English learner status and student mobility.

Unfortunately, the accountability process does not produce public results until late in the summer before the next school year is to begin. By that time most school staffing decisions have likely been made. It therefore seems likely that if adjustments are to be made by the districts, or the teachers, their first opportunity will be the summer after the accountability failure is announced (A few preliminary regressions bear this out). Thus I examine the impact of failure in the 2002-2003 school year as measured by spring 2003 tests and announced in late summer 2003, on teacher experience measured in the fall of 2004 for the 2004-2005 school year, and so forth. This process allows me to examine three years of potential AYP failure (fall 2002-2004) on teacher outcomes in the falls of 2004-2006.

An OLS estimation of equation (1) is unlikely to uncover a causal relationship between failure and subsequent teacher experience for a couple of reasons. Even conditioning on prior test scores, failing schools likely differ from those that do not fail. These unobservable characteristics bias the OLS estimates. Second, even if there are no commonly understood characteristics that differ between the schools, those that fail could differ due to a barking dog effect. If those that fail had particularly bad test scores due to random error, those scores might be expected to revert to the mean in the following period. If teachers anticipate this they might not react to the failure, which OLS would mistakenly treat as a zero failure result.

I deal with this problem in two ways. First, I adopt an instrumental variables approach. Consider a relationship predicting school failure of the form:

$$F_{it-1} = \delta Z_{t-1} + X'_{it}\theta + \omega_{it-1} \qquad (2)$$

Indexing and variables remain as before. Here Z describes an additional factor that influences failure probabilities. If the parameter estimate of δ is non-zero, and if Z can properly be excluded from equation (1) describing the relationship of failure and future teacher experience, then a two-stage least squares estimation of equation (1), with Z as an excluded instrument, produces consistent estimates of φ, the parameter of interest. In this case equation (2) describes the first stage relationship, which is explored in the next sub-section.

Although this approach is similar in concept to the regression discontinuity design discussed by Campbell (1969) that has recently been popular in the education literature (e.g. Angrist and Lavy 1999; Guryan 2001; Jacob and Lefgren 2004), there is one major difference. In the regression discontinuity approach the variable of interest becomes the index variable and observations with small differences in that index are considered likely to have only trivial differences in other, often unobserved characteristics. Thus large, discontinuous changes in outcomes over a small range of the index can often be attributed to the program or policy in question. However in the present situation attempts to compare schools just below a failure cutoff along an achievement index with those just above are difficult because those passing schools likely responded differently to accountability incentives.

In my approach, the non-linearity does not involve the previous test score that produced a school failure at all. Instead, a small change in the racial or economic composition of a school leads to a large change in the probability of failure. Thus instruments for failure can be constructed based on the schools' previous subgroups, controlling for demographic characteristics.

Additionally, because the subgroups are based on the racial composition of the school, I restrict the sample for analysis to schools within a small window of variability in that racial group. This helps to ensure that the schools are comparable except for the differences in failure probabilities brought about by small demographic changes resulting in an extra subgroup.

V. Results

A. Graphical Evidence

Figure 1 shows the raw relationship between the number of subgroups in a school and the fraction of those schools that fail to meet the AYP standard. The positive relationship suggests that extra subgroups generally increase failure probabilities. Figure 2, furthermore, shows that schools with different numbers of subgroups have large differences in levels of teacher experience. However, this interpretation is likely overly simplistic. First, some of the possible subgroups have very few schools that qualify. For example there are only a handful of schools with Pacific Islander or Filipino subgroups..

Second, it is possible that these graphs merely reflect other differences in schools with more racial heterogeneity that have nothing to do with subgroups.

In order to address this possibility, I will look at how the presence or absence of a particular subgroup designation within a limited window of racial variability affects failure probabilities and teacher outcomes. Figure 3 plots the underlying relationship between the percentage of Hispanic students at a school and the fraction of schools that have a significant Hispanic subgroup. While less than one-fifth of the schools with 10-14% Hispanic students have a significant Hispanic subgroup the fraction increases rapidly thereafter. At 16% Hispanic students over half the schools have a significant Hispanic subgroup and among schools with 19-20% Hispanic students four-fifths have a subgroup.

This relationship is almost certainly due to the nature of the cutoff rule. If the cutoff were an absolute number, say 15 percent, we would expect to see a discontinuous jump at exactly 15 percent as schools become eligible for that subgroup and their failure probability increases. However, in reality the numerically significant designation depends upon the number of students in the group and the entire number in the school. The fifteen percent rule only binds when there are at least 50 students in the subgroup, and the percent rule does not matter if there are 100 students in the subgroup. Thus instead of a sharp introduction of subgroup eligibility it appears to phase in as schools pass between 10 and 20 percent of Hispanic students. In fact the data shows that this is actually the region when subgroup eligibility phases in. Figure 4 shows a similar pattern for black students. That these subgroup changes result in higher failure probabilities can be seen in Figures 5 (for Hispanics) and 6 (for blacks).

B. The relationship between subgroups and failure.

A more systematic analysis of the relationship between subgroups and failure probabilities is given in Tables 2-3, which present coefficient estimates from a number of different regressions following the general form of equation (2). For a school that is trying to meet a performance goal, the presence of a subgroup can be harmful in a couple of ways. First, suppose that the true quality of all subgroups is the same as the school as a whole. Then if we consider testing as drawing a random component to combine with the true quality, meeting the target requires a draw from the random distribution above a specific cutoff. If the school adds a subgroup, it now needs to have more than one draw above that cutoff, which will always have lower probability. Second, as pointed out by Kane and Staiger (2002), the nature of a subgroup is to contain a relatively small number of students, leading to more statistical uncertainty for the subgroup and more chance of the subgroup being labeled failing. Finally, many subgroups will on average have lower scores than the school as a whole, leading to an increased failure probability.

Table 2 uses the whole sample and looks that the effects of particular indicator variables for the number of subgroups, as well as variables indicating the presence of a particular subgroup. The regressions control for student socioeconomic status including cubic controls for race. Regressions are weighted by teacher FTE at the school and standard errors reflect school level clustering. Each lettered section of the Table represents a separate set of regressions with a different set of potential instruments.

Panel A considers a set of indicator variables for the number of subgroups. Perhaps surprisingly, the presence of a second subgroup shows no statistically significant effect on failure probabilities (the point estimate is negative) once regression controls are accounted for. This can be accounted for by the fact that the coefficient reflects a mix of all the scenarios in which a second group is added. In fact it is almost as likely to be an asian subgroup added to a white majority since in many cases the presence of a hispanic or black subgroup also means there is a socioeconomically disadvantaged subgroup as well (meaning at least 3 subgroups). Indeed, for higher numbers of subgroups the relationship with failure is statistically significant and monotonically increasing.

Panel B focuses on the two subgroups that show the strongest statistical connection to failure probabilities, Hispanic and Black. The presence of either subgroup is associated with a large, significant increase in the probability of school failure to meet the performance standard. Despite the multi-collinear nature of the two subgroups they are individually and jointly significant predictors of school accountability failure.

Panel C explores an alternate parameterization of subgroup status. It considers a simple cutoff rule where an indicator variable is set equal to 1 when the fraction of school students from a subgroup exceeds 0.15. As figures (3)-(4) suggest, this is an oversimplification of reality, as subgroup status phases in over a range of minority percentages. Thus, the estimated coefficients are both smaller and less statistically significant as predictors of failure.

Table 3, considers the first stage relationship between subgroup status and failure in the two subsamples. As mentioned previously these are chosen to capture the fade-in regions for two racial subgroups while maintaining comparability among the included schools. The results show that there is a strong significant relationship between actual subgroup status as well as a cutoff indicator for racial composition >15 % and failure probabilities for a school. Having a black subgroup increases failure probability by eighteen percentage points over a similar school with no subgroup. Having a Hispanic subgroup increases the probability almost ten percentage points.

C. *The relationship between failure and future teacher experience.*

One of the primary ways in which an accountability system such as the Adequate Yearly Progress requirements of NCLB can influence a school is by providing incentives for the school not to fail in the first place. Unfortunately, this phenomenon is beyond the scope of this paper. However, we can test the hypothesis that failure might lead to a loss of teacher experience.

Table 4 presents OLS estimates of equation (1) as benchmarks. Panel A suggests that accountability failure has a negative, statistically significant effect on the level of aggregate school level teacher experience. This appears to be true across a wide variety of measures, as both average absolute experience and average district experience decline by at least one-third of a year. Given an average school size of about 30 teachers this would suggest an aggregate experience loss of about 11.5 years of total teaching experience from the school due to failure. The fraction of novice teachers appears to increase by about 2.7 percent. Additionally the fraction of teachers holding an advanced degree declines.

Because the sanctions for failure are progressive, panel B looks at a model that considers the joint effects of failing in years t-1 and t-2. Two interesting results stand out. First, the most recent year's failure remains an important predictor of teacher experience across all my measures, with some attenuation. Secondly, the effects of failure on teacher experience appear to persist after an additional year's time, and might be even larger in the future. Although I have to reduce the sample size by a year's worth of observations to produce the second lag, the results remain statistically significant.

Because the OLS results might be biased by the omission of relevant covariates, I present instrumental variables estimates of the failure to teacher experience relationship in Table 5. The sample consists of all schools with 10-20% Hispanic students.[1] The first row shows that OLS estimation on this sample presents comparable estimates to those found for the entire sample in Table 4, with the exception of district experience, which is more sharply affected by failure. The second row reports the coefficient on lagged failure when it is instrumented with Hispanic subgroup status. Across all four measures of teacher experience, the results are much larger (in absolute value), suggesting that failure due to the presence of Hispanic subgroup produces a large drop in teacher experience and increase in the fraction of novice teachers at the school. These results are not sensitive to the inclusion of socioeconomic controls (Row C) or the inclusion of quadratic and cubic terms to control for the small racial differences that do exist in this sample (row D). Furthermore, using the strict cutoff line of 15% Hispanic as an instrument instead of actual subgroup status produces a similar pattern or results, although they have reduced precision. In general these results reinforce the interpretation that under the NCLB regime failure had significant effects on future levels of teacher experience. Part of this effect is likely driven by the larger fraction of novice teachers suggested by column (3), several of which do not enjoy full accreditation status (Column (4)).

[1] Modifying the sample to include all observations for a school that falls into this window in any year produce no meaningful changes in the analysis.

Table 6 shows that this pattern of results is not due to some peculiarity of the Hispanic subsample. Indeed, the first two rows show that a similar exercise using a sample of schools with 10-20% black students produces similar estimates, with aggregate teacher experience reduced and novice teachers increasing as a result of school failure. The final two columns present a different way of viewing the results by examining windows of schools defined by the number of students of a particular race as opposed to the percentage. These windows are chosen to surround the entry point into subgroup status, 50 students. Subgroup status is then used as an instrument for failure to estimate its effect on teacher experience. Because these windows contain fewer school level observations than my earlier samples, they produce estimates with less precision. However, the pattern of point estimates for the effect of failure on teacher experience remains substantively the same.

D. Discussion

In all cases the instrumental variables estimates are larger (in an absolute value sense) than their OLS counterparts. There are a couple possible explanations. First, if the instruments work correctly, these estimates represent not the average response of teachers at failing schools, but the average across teachers at schools whose failure is due to the presence of an additional subgroup. If we interpret the presence of an additional subgroup as a sort of structural disadvantage for the school relative to the accountability criteria, then teachers who perceive that structural disadvantage may be more likely to move than teachers who perceive their schools failure as due to random events. This would imply stronger reductions in experience for these groups, which is what we observe.

Furthermore, the estimates are based on the response of schools in a narrow range of racial profiles, and the response of teachers outside this range might be different.

Second, the consistency of instrumental variables requires not only a statistically meaningful first stage relationship, but also the excludability of the instrument from the second stage. In the present circumstance this is tantamount to the assumption that the only way in which having a racial subgroup this year affects future teacher experience, once I condition on racial composition, is through its effect on failure to meet the standard (i.e. this year's test score). While this may seem reasonable, it is quite possible that the presence of an additional subgroup has effects on the desirability of a teaching job even if the school makes adequate yearly progress. For example the mere threat of failure or the additional work created for the teachers to make sure the subgroup makes adequate progress may serve as disincentives to stay at that job. If this is the case then the policy relevant effect is the direct effect of having an additional subgroup on teacher experience, ceteris paribus.

Table 7 attempts to quantify this direct effect by regressing measures of teacher experience on lagged measures of subgroup status, within some of the previously defined subsamples. Although the pattern of results looks familiar, the presence of an additional subgroup leads to significantly lower aggregate experience and more novice teachers that comparable schools, the magnitudes are smaller as the reductions occur on average across all schools with a subgroup, not just those that fail. Indeed these estimates imply that subgroups status will lead to schools losing about 8 years of aggregate experience from their teachers and gaining on average 0.3 new novice teachers.

It is also important to note that measures of teacher experience do not directly relate to student achievement, though at least one other study has suggested that failure under NCLB can also reduce student test scores (Sims 2007). In fact it could be that teachers are not leaving because of stigma or bureaucratic hassle, but because they are uninterested in improving teacher quality. Similarly, it is possible that the reduction in teacher experience is the product of choices by administrators trying to induce teacher switches in a way that would improve student achievement, rather than teacher flight. In any case, it is certainly remarkable to find some evidence that the NCLB incentive program, which is designed to put a qualified teacher in every classroom, might actually reduce the experience of teachers in at risk schools. Such an unintended consequence would certainly be worthy of note.

VI. Conclusion

The current federal accountability system for public elementary and secondary schools, as well as the systems of many states depend at their heart on the incentives embodied in the threat of failure, whether these are administrative punishments that raise the cost or threaten the jobs of personnel, the social sanction of stigma, or the potential for school choice. However, by its very nature such punishments may drive away teachers that are essential to making schools better. Though this is difficult to measure directly, this study looks at one indicator of teacher quality that is easily measurable, experience. By using racial subgroup rules as instruments I hope to capture the true effect of failure on the school level experience of teachers, as distinct from the effect of a school with low test scores, as the latter suggests important things about the school that should affect future teacher experience regardless of the accountability system.

My results suggest that subgroup rules create a system where schools with additional subgroups are much more likely to fail all other things equal. In turn, I find that schools that fail to make AYP due to the presence of additional subgroups suffer large losses in aggregate teacher experience in future years. Clearly these findings are a starting point and further research is necessary to test the robustness of this effect and explore its implications. However, if the relationship is true it is suggestive that there are potential internal limitations to the ability of high powered incentive schemes to produce the results promised by NCLB advocates.

References

Angrist, Joshua D. and Victor Lavy (1999) Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics* 114, 533-575.

Ballou, D., & Springer, M. G. (2008). Achievement trade-offs and no child left behind. Manuscript. Peabody College of Vanderbilt University.

California Department of Education. *Standardized Testing and Reporting Data*. Years 1998-2000.

______, PAIF Files. Years 2003-2006.

______, Adequate yearly progress record data files. Years 2002-2005.

California State Legislature. Public Schools Accountability Act of 1999. Downloaded online from http://www.leginfo.ca.gov/pub/99-00/bill/sen/sb_00010050/sbx1_1_bill_19990405_chaptered.html

Campbell, Donald T. (1969). Reforms as Experiements. *American Psychologist* 24, 409-429.

Carnoy, M., Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Education Evaluation and Policy Analysis 24*, 305– 331.

Figlio, David N., and Cecilia Elena Rouse. (2006). Do accountability and voucher threats improve low-performing schools? Journal of Public Economics 90, 239-255.

Greene, J., 2001. An Evaluation of the Florida A-Plus Accountability and School Choice Program. Working Paper. Manhattan Institute for Policy Research.

Goldhaber, Dan D., and Dominic J. Brewer. (1997). Why don't school and teachers seem to matter? *Journal of Human Resources* 32, no. 3:505–23.

Guryan, Jonathan. (2001). Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts. *NBER Working Paper #8269*.

Hanushek, Eric A. (1997). Assessing the effects of school resources on student performance: An update. *Education Evaluation and Policy Analysis* 19:141–64

Jacob, Brian, and Steven Levitt. (2003) Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics 118*, 843-77.

Jacob, B. and Lefgren, L. (2004). "The Impact of Teacher Training on Student Achievement: Quasi-ExperimentalEvidence from School Reform Efforts in Chicago." *Journal of Human Resources 39*, 50-79

Kane, Thomas J., and Douglas Staiger. (2003). "Unintended Consequences of Racial Subgroup Rules" in Paul E. Peterson and Martin R. West (eds.) No Child Left Behind? The Politics and Practice of Accountability. Washington, DC: Brookings Institution Press.

Kane, Thomas J., and Douglas Staiger. (2002). "The Promise and Pitfalls of Using Imprecise School Accountability Measures" *Journal of Economic Perspectives 16*, 91-114

Ladd, H., Glennie, E., (2001). A replication of Jay Greene's voucher effect study using North Carolina data. In: Carnoy, M. (ed.), Do School vouchers Improve Student Performance? Washington, DC: Economic Policy Institute.

Neal, D., & Schanzenbach, D. W. (Forthcoming). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*.

Sims, David P. (2008). Strategic responses to school accountability measures: It's all in the timing. *Economics of Education Review* 27 (1), 58-68.

Sims, David P. (2007) Can failure succeed? Using racial subgroup rules to analyze vouchers, stigma, and school accountability. Mimeo. Brigham Young University.

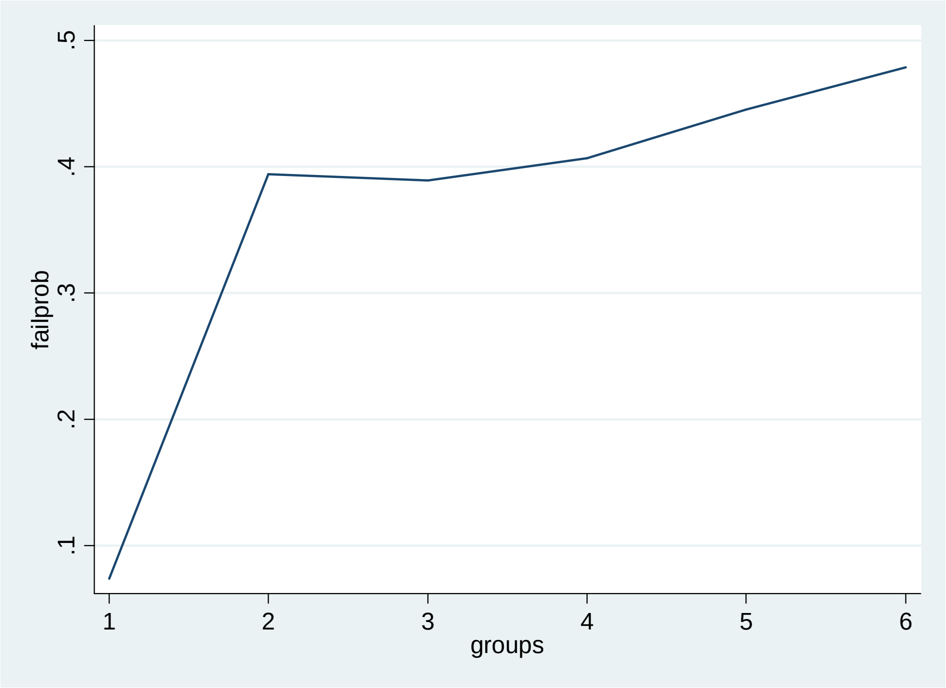

Figure 1: Relationship between number of subgroups and failing NCLB standard in California 2003-2005.

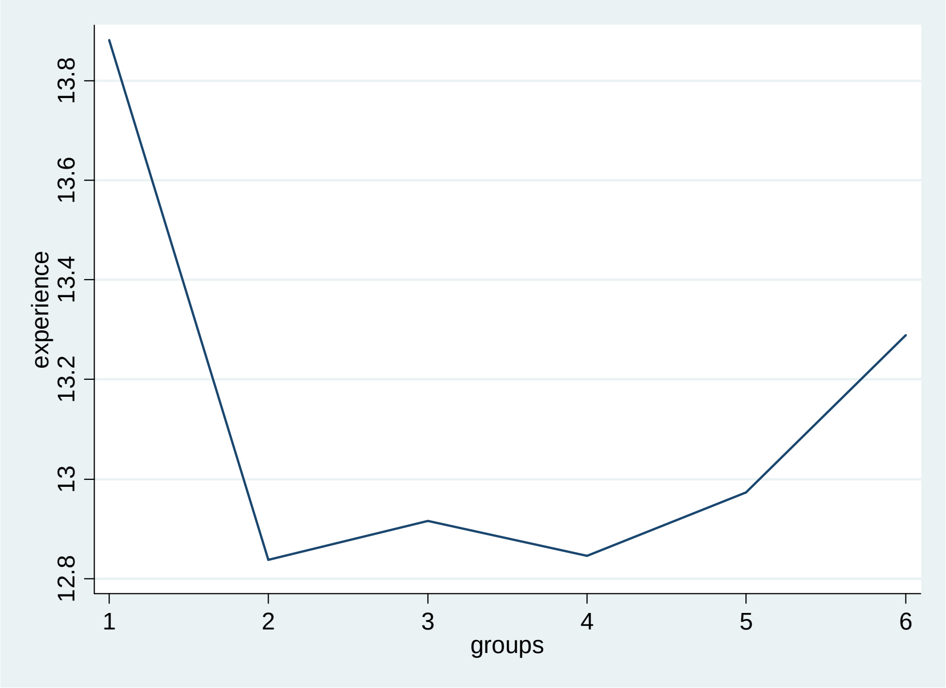

Figure 2: Relationship between number of subgroups and years of teacher experience in California 2003-2005.

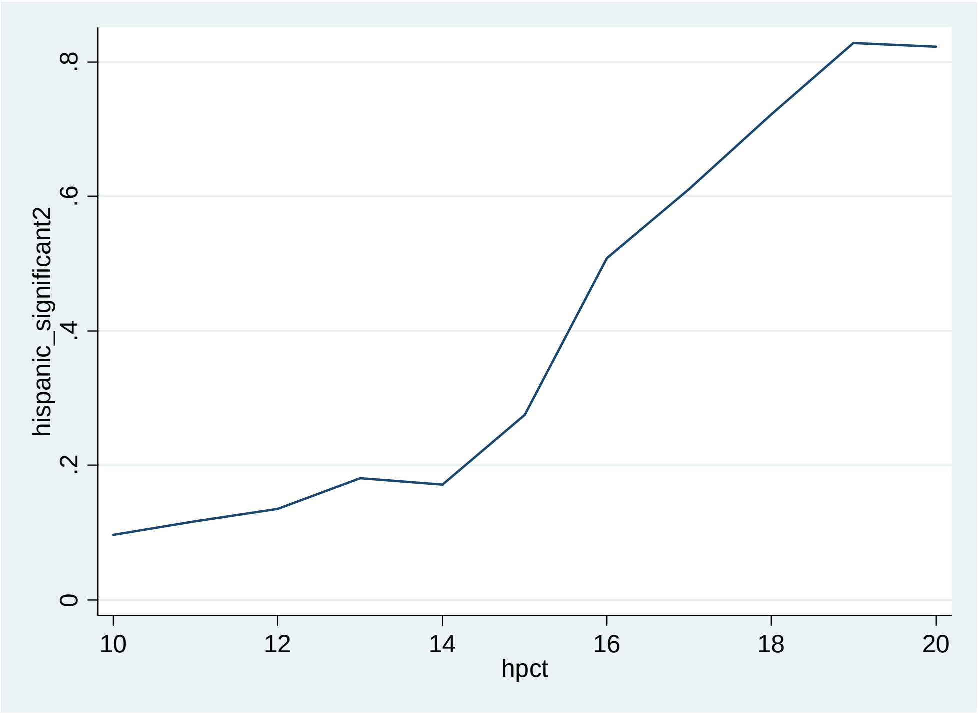

Figure 3: Relationship between Hispanic student percentage and Hispanic subgroup status

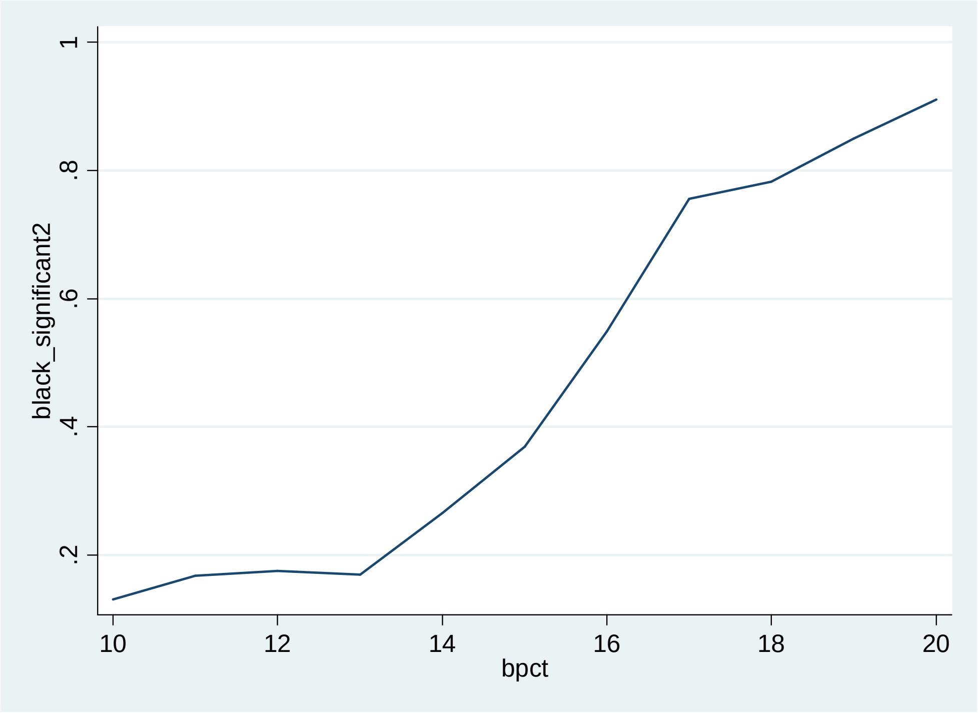

Figure 4: Relationship between Black student percentage and Black subgroup status

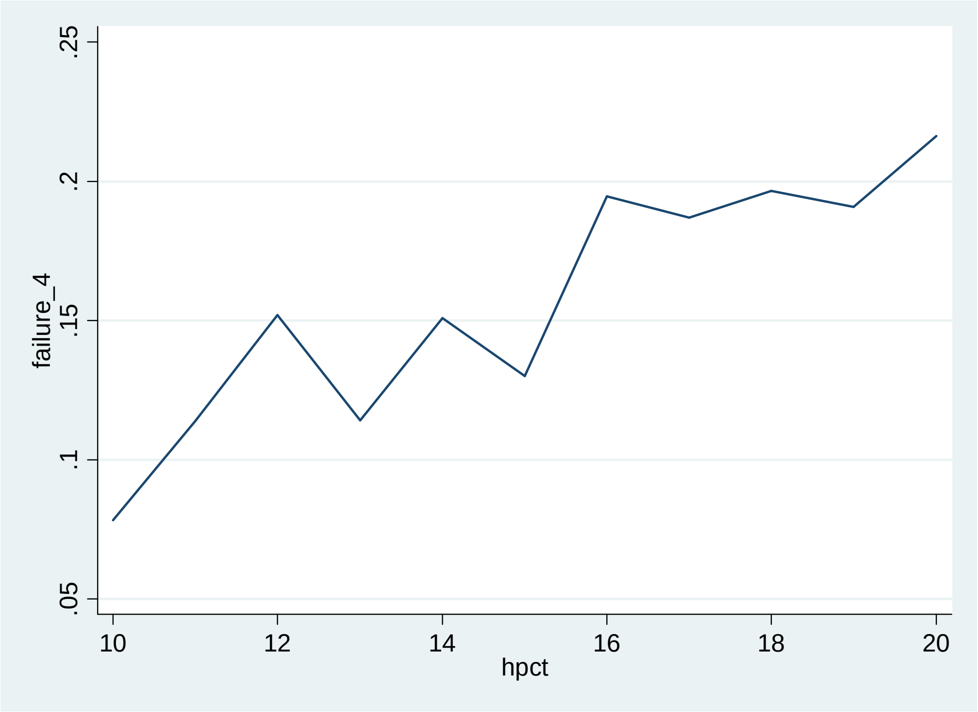

Figure 5: Relationship between Hispanic student percentage and failing NCLB standard

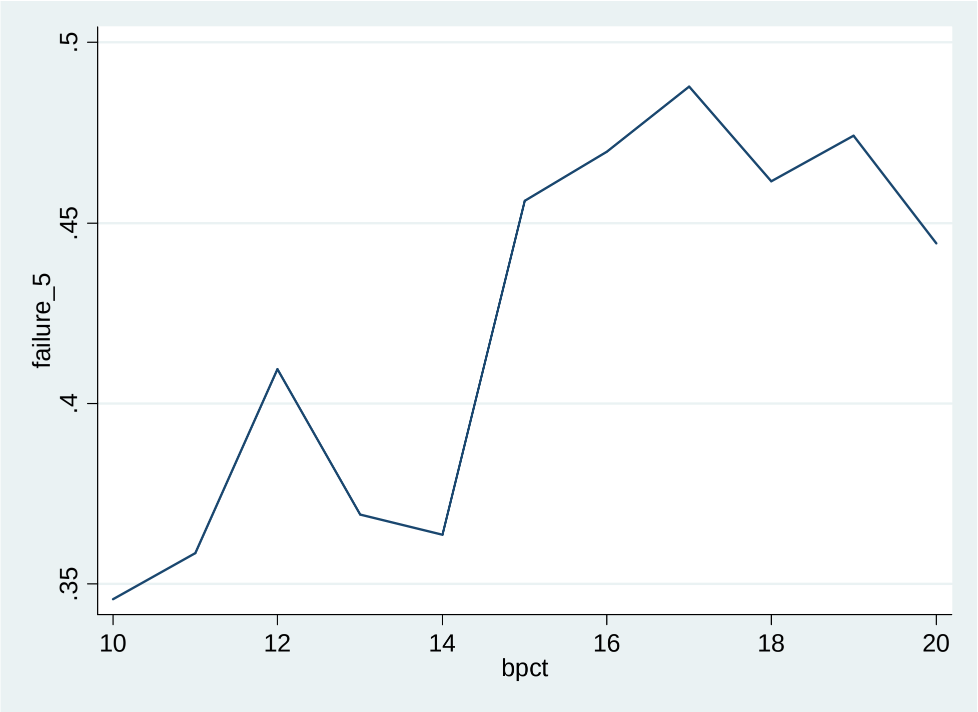

Figure 6: Relationship between Black student percentage and failing NCLB standard

Table 1 – School Descriptive Statistics

Sample	All	10-20% hispanic	10-20% black
Variable			
Teacher Experience	12.983	13.964	12.558
	(3.453)	(3.528)	(3.306)
District Experience	10.782	11.408	10.616
	(3.199)	(3.419)	(3.201)
Fraction Novice	0.162	0.142	0.169
	(0.118)	(0.111)	(0.113)
Fraction emergency credential	0.028	0.021	0.032
	(0.053)	(0.045)	(0.052)
School failed at time (t-1)	0.363	0.153	0.410
	(0.481)	(0.360)	(0.492)
Fraction Hispanic Students	0.433	0.147	0.444
	(0.291)	(0.032)	(0.212)
Fraction Black Students	0.078	0.080	0.142
	(0.114)	(0.134)	(0.032)
Fraction Disadvantaged Students	0.530	0.266	0.613
	(0.313)	(0.209)	(0.259)
Fraction Free Lunch	0.503	0.256	0.584
	(0.308)	(0.207)	(0.260)
Fraction English Learner	0.254	0.088	0.258
	(0.219)	(0.089)	(0.181)
Mobility	0.175	0.169	0.207
	(0.113)	(0.122)	(0.118)
Significant subgroups	2.637	2.368	3.177
	(0.940)	(1.111)	(0.969)
Hispanic Subgroup Present	0.759	0.384	0.894
	(0.428)	(0.486)	(0.308)
Black Subgroup Present	0.164	0.138	0.396
	(0.370)	(0.345)	(0.489)
n	21009	3660	3011

Standard deviations are given in parentheses below means. The second column contains a sample of all school year observations with between 10 and 20 percent of the student body composed of Hispanic students. The third column contains a sample of all school year observations with between 10 and 20 percent of the student body composed of black students.

Table 2: Relationship between subgroups and failure to meet NCLB standard.

A. Subgroup indicators	Whole sample (n=21009)
2 subgroups	-0.008
	(0.012)
3 subgroups	0.092*
	(0.015)
4 subgroups	0.143*
	(0.018)
5 subgroups	0.169*
	(0.030)
6 subgroups	0.211*
	(0.075)
F-stat	19.69
B. Common subgroups	
Hispanic subgroup	0.090*
	(0.016)
Black subgroup	0.139*
	(0.022)
F-stat	39.58
C. Racial Cutoffs at 15%	
Hispanic cutoff	0.038*
	(0.018)
Disadvantaged cutoff	0.051*
	(0.025)
F-stat	3.94

Each panel of reported coefficients represents a separate linear probability regression with the outcome =1 if the school failed to meet AYP and = 0 otherwise. Standard Errors corrected for clustering at the school level in parentheses. *- significant at a %5 level. All regressions control for average socioeconomic characteristics of students as well as class size and student mobility rates.

Table 3: Relationship between subgroups and failure in racial subsamples.

Sample	Variable	
Schools with 10%-20% Hispanic students (n=3660)	Has a Hispanic Subgroup	0.098** (0.012)
	Has >15% Hispanic	0.064** (0.012)
Schools with 10%-20% Black students (n=3011)	Has a Black Subgroup	0.180** (0.018)
	Has >15% Black	0.098** (0.017)

Each panel of reported coefficients represents a separate linear probability regression with the outcome =1 if the school failed to meet AYP and = 0 otherwise. *- significant at a 5% level

Table 4: Least Squares estimates of effect of NCLB accountability failure on school level teacher characteristics

Teacher Characteristic:	Teacher Experience	District Experience	Fraction Novice	Fraction emergency credential
	(1)	(2)	(3)	(4)
A. Single failure lag				
School failed at time (t-1)	-0.384*	-0.551*	0.027*	0.017*
	(0.064)	(0.060)	(0.002)	(0.001)
B. Multiple failure lags				
School failed at time (t-1)	-0.303*	-0.356*	0.018*	0.011*
	(0.069)	(0.063)	(0.003)	(0.002)
School failed at time (t-2)	-0.380*	-0.534*	0.024*	0.016*
	(0.068)	(0.063)	(0.002)	(0.002)

Standard Errors corrected for clustering at the school level in parentheses. *- significant at a %5 level. All regressions control for average socioeconomic characteristics of students as well as student mobility rates. Regressions are weighted by school FTE. N=21,009.

Table 5: The relationship between failure to meet NCLB standard and future teacher characteristics – Hispanic 10%-20% sample.

Teacher Characteristic:	Teacher Experience	District Experience	Fraction Novice	Fraction emergency credential
Method	(1)	(2)	(3)	(4)
A. OLS on this sample	-0.359*	-0.833*	0.027*	0.017*
	(0.143)	(0.137)	(0.005)	(0.002)
B. Instrumental Variables	-2.749*	-2.664*	0.104*	0.056*
- does school have a Hispanic subgroup?	(1.254)	(1.203)	(0.039)	(0.016)
C. + demographic controls	-2.629*	-3.111*	0.116*	0.065*
	(1.261)	(1.228)	(0.040)	(0.017)
D. +higher order race terms	-3.395*	-3.557*	0.099*	0.074*
	(1.607)	(1.552)	(0.049)	(0.021)
E. Instrumental variables – Indicator for >15%	-2.387 (1.859)	-1.576 (1.769)	0.086 (0.058)	0.024 (0.023)

Standard Errors corrected for clustering at the school level in parentheses. *- significant at a %5 level. Demographic controls add mobility and student socioeconomic measures. Higher order race terms include quadratic and cubic. Regressions are weighted by school FTE. N=3660.

Table 6: The relationship between failure to meet NCLB standard and future teacher characteristics – other samples.

Teacher Characteristic:		Teacher Experience	District Experience	Fraction Novice	Fraction emergency credential
Sample	Instrument	(1)	(2)	(3)	(4)
Black 10-20% sample (n=3011)	Black subgroup	-1.846* (0.677)	-2.783* (0.671)	0.109* (0.024)	0.084* (0.012)
	Black students >15%	-4.917* (1.383)	-6.120* (1.470)	0.0141* (0.044)	0.039* (0.019)
Schools with 40-60 hispanic students (n=1587)	Hispanic subgroup	-3.903 (2.520)	-2.312 (2.331)	0.332* (0.101)	0.070* (0.035)
Schools with 40-60 black students (n=1848)	Black subgroup	-2.154 (2.036)	-3.717 (2.103)	0.041 (0.066)	0.051* (0.025)

Table 7: The relationship between subgroup status and future teacher characteristics.

Teacher Characteristic:	Teacher Experience	District Experience	Fraction Novice	Fraction emergency credential
Sample	(1)	(2)	(3)	(4)
A. Hispanic 10-20% sample	-0.270*	-0.262*	0.010*	0.005*
(n=3660)	(0.120)	(0.116)	(0.004)	(0.002)
B. Black 10-20% sample	-0.331*	-0.500*	0.019*	0.015*
(n=3011)	(0.123)	(0.119)	(0.004)	(0.002)
C. Schools with 40-60 hispanic	-0.316	-0.186	0.027*	0.006*
students. (n=1587)	(0.202)	(.187)	(0.006)	(0.002)

CALDER