Trick or treat: new ecology of education accountability system in the USA

Jaekyung Lee*

SUNY at Buffalo, Buffalo, USA

(Received 15 June 2009; final version received 13 September 2009)

This study tracks American states’ policy choices under the No Child Left Behind Act and explores their consequences for student achievement. Using the path analysis of relationships among state-level policy input, context, and outcome variables, the study portrays a Halloween-like ‘trick-or-treating’ game between the federal and state governments in the new ecology of the test-driven education accountability system. States that chose the ‘trick’ path with a calculative policy negotiation and manipulation strategy made significant gains on their own state assessments but not on the national assessment. In contrast, states that followed the ‘treat’ path with a faithful policy implementation for funding strategy have not yet brought about significant gains on either the national or state assessments. The first-generation accountability states with a prior history of high-stakes testing tended to employ both strategies at the same time. However, neither effective illusion nor ineffective implementation serves the goal of long-term, sustainable academic improvement. Implications for research and policy are discussed.

Keywords: high-stakes testing; accountability; NCLB; assessment; federalism; policy analysis

The No Child Left Behind Act of 2001 in the USA is aimed at ensuring 100% reading and math proficiency for all students across all states by 2014 (NCLB 2001). While NCLB allows individual states to set adequate yearly progress (AYP) targets based on states’ own performance standards and assessments, states have to monitor and report the percentage of schools that did not make AYP and the percentage of schools identified as in need of improvement. Initially, it was estimated that up to 80% of schools in some states could be targeted as needing improvement or corrective action in the first few years under NCLB (Lee 2004; Marion et al. 2002; Olson 2002). According to several national and state reports of actual AYP statistics in the past six years of NCLB implementation, it turns out that the percentage of schools identified as being in need of improvement, at least in the first several years, was of a much smaller fraction than expected, despite variations among the 50 states and signs of increasing rates (CCSSO 2008; Education Week 2004, 2006; LeFloch et al. 2007; Padilla et al. 2006): 13% of all schools rated for AYP in 2002–2003 through 2004–2005, 17% in 2005–2006, and 21% in 2007–2008.1

While massive school AYP failure did not occur, there were instead many state reports of academic progress on their own assessment measures of student achievement.
in reading and math. A quick glance at the aggregate statistics shows that, on average, states made some progress in terms of student proficiency rates and school AYP pass rates after the first three years of NCLB implementation. For proficiency rates, states gained about 6% on average across grades in reading and math (ranging from −7% to 28% gain) between 2003 and 2005. At the same time, states also gained about 8% in AYP pass rate (ranging from −18% to 39% gain). Despite the overall progress, there exist substantial interstate variations in these indicators of academic progress, which may be related to interstate variations in state policy responses to NCLB.

Interstate variation in student achievement outcomes might be attributable to two possibilities. One possible reason for the variation might be states’ differential implementation of the law that rests on its presumption about strong efficacy of high-stakes accountability. NCLB builds upon the alleged success of first-generation accountability states such as Texas, which adopted test-driven educational accountability systems prior to NCLB. However, the evidence of the effects of high-stakes testing was mixed and often contradictory (see Carnoy, Loeb, and Smith 2001; Grissmer and Flanagan 1998; Grissmer et al. 2000; Haney 2000; Hanushek and Raymond 2004; Ladd 1999; Phelps 2005; Skrla et al. 2004; Valencia et al. 2004). A meta-analysis of studies on pre-NCLB high-stakes accountability policy effects reveals uncertainties and inconsistencies of the findings (Lee 2008). Further, recent studies on post-NCLB academic progress give mixed findings (see Center on Education Policy 2006, 2007; Duffett, Farkas, and Loveless 2008; Education Trust 2004, 2006; Fuller et al. 2006; Lee 2006; Zimmer et al. 2007). Previous studies showed that test-driven accountability policies imposed changes in schools with little to no support over the long haul and that this unfunded mandate has created problems (Kim and Sunderman 2004; Levin 2001; Linn 2003; Porter and Chester 2002; Sunderman, Kim, and Orfield 2005; Smith, Heinecke, and Noble 1999). The other possible reason for the interstate variation in policy outcomes is states’ use of differential standards and test measures for student and school accountability. Many states developed their own policy instruments including standards and tests under the current dual federal–state education accountability system. Given the tradition that much of educational policymaking has been historically left to the states, states were able to negotiate an implementation plan with the federal government and capitalize on built-in flexibilities in federal rules and regulations regarding standards under NCLB (Mills 2008). This room for manipulation and accommodation allowed for setting and adjusting the level of statewide student proficiency standards and/or school AYP standards in such ways to make their student and school performance look better on report cards. Under NCLB, state and local education agencies (LEA) could have learned tricks not only on how to comply with a complex set of federal rules and regulations for policy implementation but also on how to meet higher expectations for test results with limited resources. Observed survival strategies include goal reduction (e.g., lowering cut scores for proficiency standard), the procrastination of policy implementation (e.g., delaying the schedule of proficiency target raises for schools), and exclusion of low-achieving students from testing or reporting, all of which threaten the validity of test-driven external accountability policy (see Davis 2006; Kim and Sunderman 2004; LeFloch et al. 2007; Popham 2004).

In light of these concerns, this study examines interstate variations in both measurable indicators of policy inputs and student outcomes. How did the states respond to the federal NCLB policy mandates and translate them into their own goals, actions, and outcomes? Is the test-driven external accountability policy working as intended to
improve student and school performance under NCLB? On which test measures of academic achievement did the states show more or less improvement, national or state assessment? The purposes of the study are to differentiate policy pathways under NCLB and to estimate or predict which path is chosen by which type of states. It will help us better understand: (1) how states responded to the test-driven external accountability policy mandates under NCLB, and (2) how their different responses generated different results on different types of assessments, and possibly different impressions of the states’ academic progress and policy success under NCLB.

New ecology of American education accountability system

The origin of NCLB can be traced back to Title I of the Elementary and Secondary Education Act of 1965, federal compensatory education program for disadvantaged students. The states’ or local Title I policy implementation has been depicted as a process of ‘mutual adaptation’ (McLaughlin 1976), ‘ecology of games (non zero-sum games)’ (Kirst 1995), and ‘blending development incentives with the redistributive requirements’ (Peterson, Rabe, and Wong 1986). The past cooperative federal–state relationship, so-called ‘marble cake’ federalism, has been replaced by more complex relationship driven by both cooperation and competition. What has emerged under NCLB may be called ‘bipolar federalism’ in that this new federal–state relationship reveals some features of both cooperative and competitive federalism at the same time. Cooperative federalism is a concept of federalism in which national, state, and local governments interact cooperatively and collectively to solve common problems (Elazar 1962). In contrast, competitive federalism is a concept of federalism in which conflict and bargaining are viewed as the keystones that perpetuate intergovernmental balance (Rosenthal and Hoefler 1989). These two concepts of federalism are not incompatible with each other in practice, as states employ multiple, diversified strategies in response to a common federal policy thrust. For example, the states’ faithful implementation of fundamental parts of the law (e.g., the NCLB mandate of testing all eligible students in Grades 3–8 in reading and math) in return for federal funding demonstrates cooperative federalism on one hand, but at the same time their engagement in calculative negotiation and manipulation of administrative rules and regulations about testing standards or even addition of their own state educational accountability policies in anticipation for better results signifies competitive federalism on the other hand.

New institutionalism theory can help provide new insights into the complexity of changing federal–state relationships in American education along with the diversity of states policy strategies and outcomes under NCLB. The new institutionalism theory suggests that school systems which depend on conformity to institutionalized myths for survival often engage in a process of decoupling that buffers work in the technical core from the consequences of institutional conformity; the process is expected to gain public confidence by maintaining the appearance that things are working as they should be, even if they aren’t (Meyer and Rowan 1978; Rowan and Miskel 1999). The performance-driven external accountability movement, such as NCLB, has brought significant changes to the traditionally decoupled environment of educational systems. With improved educational technology and enhanced capacity to inspect educational productivity, school systems face a much stronger demand for technical performance improvement, but at the same time they do not experience a decline in demands for institutional conformity (Rowan and Miskel 1999). Given the high-stakes nature of
these multiple demands, state governments are likely to choose multiple adaptation strategies for maximizing the chance of their institutional survival. In other words, state policymakers become concerned about both legitimacy and productivity, not only adopting practices that they believe the federal government deems appropriate or legitimate but also seeking measurable results on the efficacy of adopted practices.

Although NCLB establishes states’ own assessments as the basis of evidence for school accountability, the federal assessment for so-called nation’s report card – National Assessment of Educational Progress (NAEP) – plays a confirmatory role as an independent assessment to validate the state test results.\(^2\) Prior research showed significant discrepancies between NAEP and state assessment results, suggesting that for many states, NAEP proficiency levels are more challenging than the states’ own (Fuller et al. 2006; Klein et al. 2000; Koretz and Barron 1998; Lee 2007; Linn, Baker, and Betebenner 2002). A recent study found that the observed heterogeneity in states’ reported proficiency rates can be largely attributed to differences in the stringency of their standards: most states’ proficiency standards fall below the NAEP Proficient standard, and many fall below the NAEP Basic standard (NCES 2007). Since state standards vary widely in their relationship to NAEP standards, questions arise about the generalizability of gains reported on a state’s own assessment, and hence about the validity of claims regarding student achievement (Linn 2000). Recent studies demonstrate that since NCLB’s inception, state assessment results show improvements in math and reading, but students are not showing similar gains on the NAEP – the only independent national test (Lee 2006). Therefore, assessing academic progress as an indicator of policy outcomes under NCLB requires cross-validation of evidence from both NAEP and state assessment results for matched student cohorts tested in the same years, subjects, and grades. Although all levels of the educational system have responsibility for implementing the provisions of NCLB, states play a particularly important role in that they adopt the standards-based policies that determine the accountability goals and interventions throughout their jurisdictions (LeFloch et al. 2007). Federal K-12 funds are sent to the state education agencies (SEA) with subsequent distribution requirements to pass the majority of the funds through to the LEA. While LEAs are responsible for determining whether individual schools make AYP thresholds, the state is ultimately responsible for ensuring that all requirements of NCLB are being met. This situation is described by state legislatures: ‘money goes to the school district whereas responsibility goes to the state’ (National Conference of State Legislatures 2003). This new policy environment raises a question as to whether states can be treated as ‘unitary’ actors under NCLB, when there exist multiple education policy-making institutions and pluralistic policy processes at the state level. During the past two decades of the standards-based education reform movement prior to NCLB, state legislatures have played a major role, even though responsibility for formulating and administering most reform policies was shared by state legislatures, state boards of education, state departments of education, and the state governor (Fuhrman 1988; Lee 1997; Mazzoni 1993). Under NCLB, it seems that state legislatures continue to have played a major role by pushing for timely implementation of federal mandates on one hand and calling for funding, fixes, and waivers on the other hand (Communities for Quality Education 2008). This study assumes that states function as unitary actors in their responses to NCLB and that a uniform set of factors constrains policy alternatives considered by states.

In the new ecology of the American education accountability system, this study attempts to identify the states’ policy paths and explore their potential consequences...
for student and school performance. Two possible policy scenarios emerge for states’ two seemingly divergent responses to NCLB, using the metaphor of Halloween trick-or-treating game. The literature review suggests that states may play a new game of ‘trick or treat’ as an institutional ritual as they attempt to figure out survival strategies under the new ecology of the test-driven educational accountability system with tighter resources and higher pressure from the federal government.

Following the enactment of the law, the federal government was expected to give states timely guidance and adequate financial support for policy implementation. If the federal government met the expectation of giving ‘treats’ to states, states would treat the feds in return through faithful and timely implementation of NCLB. On the other hand, if the states felt undue federal and public pressure for meeting the goal of 100% proficiency without timely and adequate support, states might choose to trick the feds and public instead. While states are allowed to set school AYP targets at their own paces like moving targets, they can also adjust the rigor of their student performance standards for proficiency. Perverse uses of these options may generate an illusion of academic progress. However, some states might have actively implemented the law even without desired support, if they had strong loyalty to the federal government or NCLB with a shared belief in the effectiveness of the test-driven external accountability policy. Particularly for the states which had fewer educational resources on their own and thus were more dependent on federal funding, noncompliance should not be an option.

Whereas the treat path is a formal policy route that involves faithful implementation of policy mandates, the trick path is an informal policy route that involves calculative negotiation, manipulation and illusion. Despite seemingly contradictory and irreconcilable paths, both trick and treat strategies may operate together at the same time among states that attempt to maximize their chance of institutional survival. The choice or mix of these strategies may depend on several factors including the states’ pre-NCLB history of activism in high-stakes accountability policy, their loyalty to the federal government policies, and the states’ own educational capacity, and resources independent of federal support. First, states are less likely to adopt relatively expensive and complex policy strategies (Firestone 1990; Peterson 1976). For this reason, the trick path can be more attractive than the treat path for states with fewer resources and capacity. Secondly, states’ responses to the federal press for education reform are likely to vary by compatibility with their own policies as embedded in political culture (Firestone, Fuhrman, and Kirst 1989; Mazzoni 1993). Therefore, comparing states in terms of their high-stakes accountability policies and outcomes requires differentiation of the first-generation and second-generation accountability states. States that adopted accountability polices before NCLB are called ‘first-generation’ accountability states and include Kentucky, Maryland, North Carolina, California, Florida, New York and Texas (Mintrop and Trujillo 2005). In contrast, states that never initiated statewide accountability reform before NCLB are called ‘second-generation’ accountability states for the sake of distinction (whether they embraced NCLB or not).

Method
This study draws on multiple sources of data to construct key policy input, context, and outcome variables for 50 states (see Appendix 1 for description of the variables). Figure 1 shows the analytical framework for this study to explore possible causes and
consequences of states’ different policy strategies under NCLB. The arrows show hypothesized relationships among variables. The study conducts path analysis with state-level aggregate variables in order to test hypotheses based on aforementioned policy scenarios. The path analytic model is chosen to help track chains of relationships between policy input, context, and outcome variables based on the new model of American educational accountability system.

As shown in Figure 1, the study attempts to track two hypothesized policy paths on student achievement outcomes: (1) ‘trick’ path (following the path of upward-moving arrows that originate from ‘state activism in pre-NCLB high-stakes test accountability’ variable in Figure 1) that presumes states’ calculative negotiation and manipulation of rules and regulations regarding testing and standards based on the model of self-interested policymaking and competitive federal–state relationship, and (2) ‘treat’ path (following the path of downward-moving arrows from the same first variable in Figure 1) that presumes states’ faithful implementation of federal policy mandates based on the model of shared policymaking and cooperative federal–state relationship.

Under NCLB, the existence of dual accountability systems and interactions between federal and state policies poses methodological challenges for the analysis of policy outcomes. It is necessary to understand variations among states in terms of their high-stakes accountability policy history prior to NCLB as well as their current state of NCLB policy implementation. First of all, pre-NCLB state activism in high-stakesAccountId

Figure 1. Path diagram of the relationships between state policy strategies and achievement outcomes (statistical significance of standardized path coefficients shown on arrows are noted by *p < .05, **p < .01, and ***p < .001; R = reading and M = math).
test-driven accountability, as a NCLB policy precursor variable, is expected to predict how states respond to federal policy change under NCLB. Particularly, this study uses the measure of state activism in high-stakes testing accountability constructed by Lee and Wong (2004).

Second, the states’ policy implementation and funding variables are measured during the period of NCLB, which would capture the hypothesized process of ‘treat’ path. The study constructs the measure of state NCLB policy implementation as crude proxy indicators of aggregated state-level policy activities based on the Education Commission of the States (ECS 2007) report. Although this information reflects only state-level policy efforts and not local-level ones, it is supposed to discriminate states in terms of how well they have complied with federal mandates since the enactment of NCLB in its key components (see Appendix 1).

Third, the states’ student proficiency and school performance standard variables are measured, which would capture the hypothesized process of ‘trick’ path. These measures related to the rigor of tests and standards are chosen because they are the most crucial variables of the federal high-stakes testing accountability system, but at the same time most vulnerable to negotiation and manipulation by the states. Specifically, the rigor of state proficiency standards is measured by the discrepancies between NAEP proficiency rates and state assessment proficiency rates. The rigor of school AYP standards is measured by annual measurable objectives (AMO).

These measures of both ‘trick’ and ‘treat’ policy processes are expected to affect student achievement results as measured by state assessments and/or NAEP. The evaluation of statewide academic improvement under NCLB has been tricky and controversial, because these different test measures often produce different results. Further, academic progress can be measured directly by gains in student proficiency rates and indirectly by gains in school AYP pass rates. Although the state definition of AYP is based primarily on the state’s academic assessments, it is determined not only by student proficiency in reading and math, but also by other non-cognitive performance indicators such as attendance and graduation rates. For measures of student achievement and school progress, this study focuses on reading and math proficiency rates for Grades 4 and 8, from both NAEP and state assessments, and school AYP pass rates during the period of 2003–2005.

These hypothesized policy effects are explored through path analysis. The path analysis method involves specifying the relationships among multiple variables in a series of multiple regression equations and portraying them graphically in a path diagram: causal relationships are depicted by straight arrows and correlations are depicted by curved arrows (Hair et al. 1998; Kline 2005). Figure 1 shows the path diagram with standardized path coefficients (i.e., beta weights = $b$) obtained from multiple regressions. For example, the path coefficient of .39 for the arrow between ‘state activism in high-stakes accountability’ and ‘fidelity of NCLB policy implementation’ shows that the direction and strength of the relationship between the two variables is positive and moderate: states that were active in test-driven accountability policy prior to NCLB tend to be on a faster track for NCLB policy implementation than states which did not have high-stakes accountability policies before NCLB.

While the study uses both state assessment proficiency rate and NAEP proficiency rate as components in defining and measuring the rigor of student proficiency standards, it also tests the hypothesis that those observed proficiency rates are affected by changes in the rigor of standards. This formulation makes it difficult to capture any true relationship between the rigor of state standards and student proficiency rates
(measured by either state assessment or NAEP). Owing to the fact that the rigor of state standards is defined as a function of its gap relative to the NAEP standard, the results may become tautological. To ameliorate this problem, the study takes a longitudinal approach to examine changes (as opposed to status) in proficiency rates over time and control for the initial status of proficiency rates. Nevertheless, causal inferences about the relationship between these standard and proficiency variables are not warranted.

This study also has other several limitations with caveats. First, the study starts with simplified metaphors about policy choices and assumes that the trick or treat behavior of state governments can be inferred from overt policy implementation or outcome indicators. However, it does neither measure nor represent the goals, motivations, and choices of individual state officials who were actually involved in complex and dynamic policy-making procedures. The path analysis relies on readily quantifiable measures of policy input, context, and outcome measures for cross-state comparisons. This statistical analysis needs to be supplemented by subsequent studies with observation, interview, and survey data that can lend validity about the volition and reasoning of state policymakers. Second, the study treats a whole state as the unit of analysis. Studying the role of different actors including state legislatures and administrations, and the inside politics of state education policymaking is out of the scope. Further, what transpires at local levels in response to federal and/or state policies is not studied. The policy measures used in this study may capture the intensity of policy enactment at the state level better than the fidelity of policy implementation at the district or school level. Third, the study does not differentiate reasons for school AYP success or failure, such as whether schools do not meet the target for all student groups or only selected racial, economic, or language subgroups. For example, 75% of the nation’s schools made AYP in 2003–2004; of the 25% that did not make AYP, half (51%) did not succeed because the school as a whole (i.e., the ‘all students’ group) or multiple student subgroups did not meet achievement standards (LeFloch et al. 2007).

Results

‘Trick’ path as an effective illusion strategy

Researchers and educators have raised concerns about unrealistically high goals and uniform performance targets for high-stakes school accountability under NCLB (Education Week 2005; Lee 2006; Linn 2003; Rothstein, Jacobsen, and Wilder 2006; Sunderman 2008). Also, education advocates, state education officials, and some members of Congress were concerned about unfunded NCLB mandates and called for more serious federal efforts to accomplish the original intent of the law (Mathis 2003; NAACP 2005; NSBA 2006). Although some argue that the cost of testing such as paying for tests and publishing test results is small relative to the cost of other expensive educational programs such as class-size reduction (Hoxby 2002), this ignores the full costs of an accountability system, which include the costs of monitoring, identifying, assisting, and doing interventions with schools in need of academic improvement. The Government Accounting Office estimated that costs associated with the NCLB testing requirement would be close to $4 billion (US GAO 2003). Although there is disagreement about the estimates of adequate funding levels, it is clear that federal support of school evaluation and technical assistance, required under NCLB,
is underfunded and the gap is likely to grow significantly with increasing identification of schools in need of improvement (Costrell and Peyser 2004).

When states do not get adequate support from the federal government and suffer from the shortfall of resources and capacity to help improve their schools’ chances of meeting AYP targets, it raises concerns about potential damage to public opinions of and support for state education system in that many public schools are likely to be identified as needing improvement and face sanctions. Given the low feasibility of meeting the ultimate goal of 100% proficiency and the high threat of losing public confidence and support for public education, underfunded mandates under NCLB may have created perverse incentives for states to game the federal accountability system. Under this scenario, one possible state policy response is reliance on an illusion strategy shown as ‘trick path’ in Figure 1. The ‘trick’ path is more likely to be chosen by states where the pressure for schools to meet high standards rise and resources remain tight. First, states may lower their performance standards in order to raise student proficiency rates. This phenomenon, dubbed ‘racing to the bottom,’ was observed among several states under NCLB (Cronin et al. 2007). Secondly, states may also attempt to control the levels of school AYP targets and hold off the schedule of raising AMOs in order to raise school AYP pass rates at least during the early phase of NCLB. Many states adopted a mixed form of AMO trajectories that project a more rapid increase in the latter years of NCLB implementation (LeFloch et al. 2007).

What accounts for interstate variation in the rigor of performance standards? One possible factor is the degree to which consequences (rewards and sanctions) are attached to state test results for schools and students. The states that have a strong accountability system with high-stakes testing would exert greater pressure for schools and students to improve their achievement on the state test than states without high-stakes accountability systems. High-stakes accountability may result in the possible inflation of the number of students reaching the proficiency level. The state education agency itself is also likely to water down its own performance standards for the sake of reducing the number of identified schools in need of improvement and intervention. The higher the stakes, the lower the state standard, and consequently the higher the student proficiency rate and school AYP pass rate. This tendency may spread across the nation under NCLB when all states are subject to the same federal requirement of high-stakes accountability. Since some states have their own accountability in addition to the federal one, the states with such dual accountability under NCLB may continue to experience greater pressure for improving test results.

Figure 1 shows that there is a significantly negative relationship between state activism in high-stakes accountability policy and the rigor of student proficiency standards in math ($b = -.07$ in reading; $b = -.27$, $p < .01$ in math). Change in the rigor of student proficiency standards is highly negatively associated with change in state assessment proficiency rates ($b = -.97$, $p < .001$ in reading; $b = -.95$, $p < .001$ in math). These results together imply that proficiency levels in states with high-stakes accountability systems may have been inflated. In contrast, change in the rigor of student proficiency standards has no bearing on NAEP reading proficiency gains, but is a significant predictor of NAEP math proficiency gains ($b = .03$ in reading; $b = .32$, $p < .01$ in math). The results indicate that the higher the states’ own performance standards relative to NAEP standards, the lower performance on state assessment (in both reading and math) and higher performance on NAEP (in math). Does NCLB high-stakes accountability drive states’ racing to the bottom? Figure 2 suggests that most states may have lowered the rigor of their student proficiency standards in reading and
math during the 2003–2005 period. Across grades (fourth and eighth grades) and subjects (reading and math), the average discrepancies between NAEP and state proficiency rates widened in a negative direction; the rigor of state standards (relative to the national standard based on NAEP proficiency level) had declined. The more states lowered their proficiency standards, the greater improvement they made in student proficiency rates on state assessments. This highly negative correlation between the two variables is explained by the fact that many states made significant gains on their own state tests without making corresponding gains on the NAEP counterparts. The correlation between NAEP proficiency gains and state assessment-based proficiency gains during 2003–2005 is close to null.

Further, the rigor of student proficiency standards is positively associated with the rigor of school AYP standards (AMO) ($b = 1.12$, $p < .01$ in reading; $b = .62$, $p < .05$ in math). At the same time, state assessment proficiency rates are positively associated with the rigor of school AYP standards ($b = 1.69$, $p < .001$ in reading; $b = 1.29$, $p < .001$ in math), which suggest that states may have adjusted the level of AYP standards according to their students’ average proficiency level. Indeed, the rigor of student proficiency standards may influence the rigor of school AYP standards in two opposite directions, canceling out the influences of each other. On one hand, states may attempt to set consistently high or low standards for both students and schools. On the other hand, states that set more rigorous performance standards for students would get relatively lower proficiency rates, which in turn results in lowering AYP standards.
The relationship between overall student proficiency rates (reading and math combined) and school AYP pass rates is positive ($b = .90, p < .01$). In contrast, the relationships between state assessment proficiency rates and NAEP proficiency rates are weak at best ($r = .20$ in reading; $r = .38, p < .01$ in math).

‘Treat’ path as an ineffective implementation strategy

Under NCLB, the working theory of test-driven accountability policy postulates that NCLB brings about significant academic improvement by giving states fiscal incentives to put test-driven school accountability system in place and holding schools accountable for test results with possible sanctions and interventions. While the primary source of NCLB implementation is the law itself (NCLB 2001), another layer of interpretation is added as the US Department of Education develops and negotiates regulations and operating guidelines for states to comply with NCLB (Mills 2008). NCLB has tightened enforcement of the law since many states did not faithfully implement its earlier version, Improving America’s Schools Act of 1994 (IASA). Under NCLB, state participation is tied to continued receipt of Title I funds, which range from about $25 million/year to about $1.2 billion/year (National Conference of State Legislatures 2003). Because exit is a much less viable option than voice for most states under NCLB, they attempted to adapt to the new policy environment by meeting high expectations with tight resources. Under this scenario, the path by which states would fully implement NCLB in accordance with federal legislative and administrative policies is labeled ‘treat path’ in Figure 1.

What accounts for interstate variation in the implementation of NCLB? On one hand, NCLB is more likely to flourish in the first-generation accountability states which share the goal of high-stakes accountability. States with strong accountability systems must have been better prepared to embrace and implement NCLB reform policy since implementation theory predicts stronger implementation fidelity among people who are accustomed to the intervention. Reforms that are isomorphic with the fundamental tenets of the institutional environment stand a better chance of survival than reforms that are not (Cuban 1992; Meyer and Rowan 1978; Rowan 1982). Lawmakers did not intend that NCLB would replace a state’s preexisting accountability policy where a parallel system already existed but rather would function as an add-on to enhance or augment state policy.

On the other hand, NCLB, if it were to work effectively as intended, is more likely to have benefited second-generation states – those states where test-driven external accountability was weak or absent prior to NCLB and where NCLB attempted to extend accountability modeled after the alleged success stories of some first-generation states such as Texas and North Carolina. By this logic, states with no exposure to high-stakes testing prior to NCLB are more likely to experience the effect of this new intervention if those states have fully implemented the law. However, there was significant threat to the fidelity of implementation, particularly in the initial phase during the first several years after the law passed. By January 2002, when NCLB took effect, only about one-third of the states had fully met the standards and assessment requirements for NCLB’s predecessor, the IASA (Erpenbach, Forte-Fast, and Potts 2003). Further, a study of NCLB implementation shows that states continue to revise content standards (American Institutes for Research 2006); since 2002, 29 states have either adopted or revised their state standards in ELA and 28 states in math. States also continue to amend their assessments and AYP plan. The study also shows variability
among states in their school support mechanism (AIR 2006): 37 states used support teams and 29 of them included individual professionals to assist schools identified as needing improvement. States tended to provide more technical assistance rather than sanctions and few schools were asked for restructuring.

The states’ average NCLB policy implementation rating as of 2007 is high: 2.76 on a scale of 1–3 (see Appendix 1 for full description of the variable). However, the implementation rating varies to some degree among 50 states in the range of 2.26–2.97 (see Figure 3). This NCLB state implementation index is positively associated with pre-NCLB state activism in test-driven accountability policy \( (b = .39, p < .01) \). The first-generation accountability states have implemented NCLB more actively than the second-generation accountability states. For example, states such as North Carolina and Texas were able to put the large-scale testing in place earlier than others in order to meet the 2005–2006 deadline that requires them to administer reading and math assessments to students in Grades 3–8. While this finding suggests that the first-generation accountability states were more ready to implement NCLB, it also implies that NCLB may not have fully spread to the second-generation states yet.

The fidelity of NCLB policy implementation is a significantly positive predictor of federal NCLB funding to states \( (b = .31, p < .01) \). This suggests that the states’ active policy implementation paid off in terms of federal funding support for their NCLB programs. However, the implementation index is not a significant predictor of state assessment proficiency gains \( (b = -.04 \text{ for reading}; b = -.02 \text{ for math}) \) or NAEP proficiency gains \( (b = -.10 \text{ for reading}; b = -.04 \text{ for math}) \). There is also a lack of

![Figure 3. Plot of NAEP 2003–2005 proficiency gains vs. NCLB implementation index.](image-url)
significant relationship between NCLB funding to states and NAEP proficiency gains in the same period ($b = -0.15$ for reading; $0.05$ for math). Figure 3 shows the scatter plot of state NCLB implementation index and NAEP proficiency gains. The results suggest that the states which implemented NCLB in more faithful and timely fashions may have received more federal funding, but they did not record more gains on NAEP and state assessments during the 2003–2005 period.7

Discussion

This study provides preliminary evidence on states’ bipolar policy responses and mixed student outcomes under NCLB. Using the metaphor of the Halloween trick-or-treating game, the study tracks two seemingly divergent policy paths and explores their consequences for reading and math achievement results based on national and state assessments in the US facing federal mandates – meeting the unrealistically high goal of 100% proficiency by 2014 – without adequate support, many states not only attempted to implement the federal policy but also outwit the policy. States that chose to ‘treat’ the federal government with a faithful policy implementation in exchange for funding strategy have not yet brought about significant gains on either the state assessment or NAEP. In contrast, states that chose to ‘trick’ the federal government with a calculative policy negotiation and manipulation strategy observed significant gains on their own state assessments in spite of the absence of corresponding gains on NAEP. The states’ choices of these different strategies can be traced back to their pre-NCLB history of high-stakes educational accountability policy. It appears that the first-generation accountability states with their own pre-existing system of high-stakes testing not only have implemented NCLB policies in more faithful and timely manners, but also pushed for quicker improvement of student test results on their own assessment with relatively lower standards.

If we consider the state education policymakers as rational, goal-oriented, value-maximizing decisionmakers working under institutional constraints as imposed by federal laws and funding formula, we can better understand states’ concurrent use of bipolar strategies – employing both a faithful policy implementation strategy in the spirit of cooperative federalism on one hand and a calculative policy bargaining and manipulation strategy in the spirit of competitive federalism on the other hand. It does not seem irrational that the first-generation accountability states, given their prior history of statewide high-stakes testing and more imminent pressure for academic improvement with tight resources, embraced the treat path as well as the trick path. In spite of the instrumental rationality of survival strategies used by those states, their long-term policy success would depend on building school capacity and resources for high-quality teaching and learning. If the USA continues the current policy course under NCLB, the Halloween-like trick-or-treating game may become an institutionalized ritual for states. More importantly, academic proficiency is unlikely to improve significantly on NAEP, and at the same time the state assessment will continue to give a false impression of progress, shortchanging children and encouraging more investment in high-stakes testing policy.

Can the USA set a new course of policy for authentic and sustainable educational improvement? There have been independent, bipartisan efforts to improve NCLB, such as the Commission on NCLB Act (Thompson and Barnes 2007), which reports to Congress and the administration with recommendations for changes based on hearings and research. Proposed changes to current NCLB accountability policies include
requiring common national standards for all states and switching from a status model to a growth model for school evaluation, which may help mitigate some problems with the trick path such as the tendency of manipulating standards and racing to the bottom. However, those proposed changes are not sufficient to address the limitations of current test-driven external accountability policy.

While the quick-fix mentality has been prevalent in the history of American education, many recurring change efforts have fallen short due to the failure to seek a more comprehensive understanding of causes for effective and sustainable solutions (Cuban 1990; Preskill and Torres 1999). It is necessary for educational policymakers and practitioners to engage in critical self-inquiry in order to understand the causes of educational problems, evaluate the intended or unintended consequences of their actions, and commit themselves to achieving the goal of authentic and sustainable academic improvement over the long run. The current cycle of school identification, sanctions and quick turnaround interventions bypass critical inquiry and capacity building stages. It is recommended that both federal and state governments coordinate educational accountability system reform efforts in critical ways: setting more realistic and differentiated goals of academic improvement, using multiple measures of school success beyond standardized test scores, providing research-based, long-term support for school capacity-building and intervention, promoting internal school accountability for continuing cycles of self-evaluation and improvement, and improving the quality and working conditions of teachers in disadvantaged, low-performing schools.

Notes
1. If schools in need of improvement continuously fail to meet their AYP targets, they are subject to subsequent stages of sanction and intervention under NCLB: (1) school transfer, (2) supplementary education service, (3) corrective action, (4) restructuring (first year), (5) restructuring (second year). The percentages of schools at each of those five intervention stages as of 2007–2008 (based on 2006–2007 testing) are relatively smaller than the percentages of schools identified for improvement, but those rates also vary widely from state to state (see Council of Chief State School Officers Profiles of state accountability systems [Online database] at http://accountability.ccsso.org).
2. The No Child Left Behind Act of 2002 mandates that states accepting Title 1 funding must participate in NAEP. Title 1 funding provides an incentive to ensure that state and local agencies provide students with the opportunity to participate in NAEP if their school is selected as part of the representative sample.
3. Standardized path coefficients are indicative of effect sizes, showing how much the dependent variable would change in standard deviation unit as a result of one standard deviation unit change in the independent variable. Based on the guidelines of effect size interpretation by Cohen (1988), the values of .10, .30, and .50 are deemed small, medium, and large effects, respectively.
4. According to the NSBA (2006), Congress has continued a steady decline in fully funding NCLB, shifting a greater portion of the cost of compliance with adequate yearly progress and supplemental services, for example, to local school districts and states. Since 2001, funding for Title I has increased by roughly 45%. However, these increases were offset by rising costs due to enrollment increases and school programming that is needed to ensure students will meet the requirements of NCLB (e.g., class-size reduction, summer school, and professional development of teachers). Further, the cost to recruit and retain highly qualified teachers and paraprofessionals, as required by NCLB, continues to grow.
5. This positive relationship reflect current federal rule about the AYP stating point. Under NCLB, states with adequate data from 2001 to 2002 were required to use those results to determine their starting points for establishing AYP targets in reading and mathematics. As required by statute, starting points were to be the higher of the percentage of students at the
proficient level in (1) the state’s lowest-achieving subgroup, or (2) the school at the 20th percentile among all schools based on enrollment, ranked by the percentage of proficient students. In most states, the percentage of students scoring at the proficient level in the school at the 20th percentile became the AYP starting point for the state (LeFloch et al. 2007).

6. There were more case where states attempted to choose the option of voice rather than exit. Numerous state legislatures have drafted resolutions to call for full funding and/or requested fixes or waivers in meeting the law’s requirements (Communities for Quality Education 2008). In some extreme cases, this led to legal actions. In 2005, the State of Connecticut sued the US Department of Education over insufficient funding and support from the federal government to help the state meet the testing provisions of NCLB (Connecticut v. Spellings). A few states attempted to opt out of NCLB without success. In 2008, the Arizona state Senate panel failed to pass a bill that would have the state opt out of the federal No Child Left Behind program (The Arizona Republic, 4 April 2008). However, the state’s withdrawal would be contingent on the Arizona Legislature backfilling whatever money the state would lose from not participating in the federal program. Since Arizona receives about $600 million a year from the federal program, the idea of NCLB opt out is not viable.

7. One may suspect that the two-year time frame is too short to capture the impact of NCLB on student achievement, if any. However, a supplementary analysis extending the time period of NAEP gain scores to four years (2003–2007) does not change the finding on the insignificance of relationships between NCLB implementation index and NAEP proficiency gains ($b = -.05$ in reading; $b = -.03$ in math).

Notes on contributor

Jaekyung Lee is an associate professor of education at the University at Buffalo, the State University of New York, and a 2009–10 fellow of the Center for Advanced Study in the behavioral sciences, Stanford University. He specializes in educational policy research and evaluation.

References


Fuller, B., K. Gesicki, E. Kang, and J. Wright. 2006. Is the No Child Left Behind Act working? The reliability of how states track achievement. PACE Working paper 06-1, University of California, Berkeley.


Appendix 1. Description of measures

State activism in pre-NCLB high-stakes test accountability

This study utilizes the measures of test-driven external accountability policy for 50 states, as constructed by Lee and Wong (2004). It is based on survey data collected in the mid- to late-1990s from three sources: (1) 1995–1996 data from the North Central Regional Education Laboratory (NCREL) and Council of Chief State School Officers (CCSSO) (NCREL/CCSSO 1996), (2) 1999 data from Quality Counts (QC) report (Education Week 1999), and (3) 1999–2000 data from the Consortium for Policy Research in Education (CPRE) report (Goertz and Duffy 2001). The NCREL/CCSSO survey covers student assessments, student accountability (testing for promotion, awards/recognition, and graduation), teacher accountability (certification gain/loss, financial rewards/penalties, probation), and school accountability (funding gain/loss, accreditation loss, awards/recognition, performance reporting, probation/warning, takeover/dissolution). The QC survey covers only student assessments and school accountability (report cards, ratings, rewards, assistance and sanctions). The CPRE survey covers student assessments and student and school accountability policies (school/district sanctions or rewards, high school exit test).

Policy index scores were calculated for each state by summing the number of policies adopted and in place by the state at the time of survey. The NCREL/CCSSO policy index ranges from 0 to 16 (M = 6.5, SD = 4.2). The reliability of this 26-item 1995 NCREL/CCSSO accountability policy index is very high (alpha = .85). The QC policy index ranges from 0 to 6 (M = 3.0, SD = 1.8). The reliability of this six-item 1999 QC accountability policy index is high (alpha = .77). Finally, the CPRE policy index was constructed by Carnoy and Loeb (2002) and the index score ranges from 0 to 5 (M = 2.1, SD = 1.4). Out of these three related policy measures, Lee and Wong (2004) created a composite factor of state activism in test-driven external accountability policy during the 1990s.

State fidelity of NCLB policy implementation

A composite index of NCLB state implementation was constructed from the ECS database which tracks state laws, departmental regulations, board rules, directives, and practices related to requirements across seven major sections of the NCLB legislation: (1) NCLB Standards and Assessments, (2) NCLB Accountability (AYP), (3) NCLB School Improvement, (4) NCLB Safe Schools, (5) NCLB Supplemental Services, (6) NCLB Report Card, and (7) NCLB Teacher Quality. This study combines the three-point scale ratings of state policy implementation status among 38 items in the seven areas (3 = on target, 2 = partially on target, 1 = not on target) as updated as of September 2007 (Available at http://nclb2.ecs.org/NCLBSURVEY/nclb.aspx?Target=AD). According to the ECS website information, the database reflects verification of enacted state policy, whether statute, rule, regulation or formal directive. The database does not include planned and implemented programs approved by the US Department as required by NCLB, but that are not yet reflected in final state policy. The goal was to
develop a ‘real-time,’ nationwide gap analysis of NCLB implementation efforts.’ For internal consistency reliability, the alpha coefficient of the composite index is .81.

**Federal NCLB funding to states**

Total federal funds for NCLB Act programs in years 2003–2005 that were allocated to states using statutory formulas. NCLB programs consist of many categories, including ESEA Title I grants to local educational agencies, school improvement grants, reading first grants, improving teacher quality state grants, and state assessments. These figures of federal NCLB funding to states were obtained from the US Department of Education website for Fiscal Year 2001–2009 State Tables: http://www.ed.gov/about/overview/budget/statetables/index.html

**Rigor of student proficiency standards**

This study measures the rigor of state reading and math proficiency standards by averaging the values of differences between NAEP-based proficiency rates and state assessment-based proficiency rates across Grades 4 and 8 in 2003 and 2005: a positive value of the gap implies a relatively higher state standard in comparison with the NAEP standard, whereas a negative value of the gap implies a relatively lower state standard; logit transformation of the gap measure was also used but the results remain highly stable. Information on 2003 and 2005 statewide reading and math proficiency rates was collected from each state’s department of education website and from the NCES website (see below). This study’s measures of the rigor of standards show concurrent validity as they have strong positive correlations with other similar measures: \( r = .97 \) for Grade 4 reading, \( r = .96 \) for Grade 8 reading, \( r = .81 \) for Grade 4 math, \( r = .85 \) for Grade 8 math in its relationships with NCES (2007) estimates of the stringency of 2005 state proficiency standard (i.e., estimated NAEP score equivalent to the state standard from Table 1) and \( r = .91 \) for Grade 4 reading, \( r = .78 \) for Grade 8 reading, \( r = .85 \) for Grade 4 math, \( r = .76 \) for Grade 8 math in their relationships with Fordham Institute (Cronin et al. 2007) estimates of the rigor of 2005 state proficiency standard. Note that the relatively weaker correlations with Fordham Institute measures may be due to some limitations: not only the smaller number of states in their study (\( N = 26 \)) but also the fact that the 2006 measure was substituted for 2005 for some states because the Fordham study reported mostly 2006 data and only less than half of the states used 2005 data instead of 2006 data.

**Rigor of school AYP standards**

This study measures the rigor of school AYP standards by each state’s AMOs. Statewide AMOs represent the percentages of students who are expected to be proficient in reading and math. AMOs vary among states and increase over time within states as moving targets (starting with baseline target in 2002 and progressing with intermediate targets toward 100% by 2014). Based on the gap between AMOs and actual proficiency rates, states determine whether schools meet performance targets each year and thus are on track toward the ultimate goal of 100% proficiency. Information on 2003 and 2005 statewide AMOs was collected from each state’s department of education website.

**State assessment reading and math proficiency**

State assessment results were obtained in the form of the percentage of students who meet state performance standard (typically at or above a Proficient level). State aggregate proficiency rates in reading and math for fourth and eighth graders in 2003 and 2005 were obtained from state department of education websites. Only fourth and eighth grade data were used in order to match them to NAEP and compare the results.
**NAEP reading and math proficiency**

NAEP assessment results were obtained in the form of the percentages of students scoring at or above Proficient level in reading and math. The data were drawn from NCES website for 2003, 2005, and 2007 NAEP public school sample Grade 4 and Grade 8 reading and math assessment results (available at www.nces.ed.gov/nationsreportcard).

**School AYP pass rate**

Percentages of schools in each state that made AYP in 2002–2003 school year were obtained from state-by-state AYP status table from Education Week report (8 December 2004). Percentages of schools in each state that made AYP in 2004–2005 school year were obtained from state-by-state AYP status table from Education Week report (20 September 2006).