Current Issues in Education
Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Volume 15, Number 3 December 1, 2012 ISSN 1099-839X

Misuse of High-Stakes Test Scores for Evaluative Purposes: Neglecting the Reality of Schools and Students

Peter Clyde Martin
Ithaca College

The article examines high-stakes test scores in Washington, DC that are used to evaluate school quality for AYP purposes. On the basis of analyses of school scores in terms of subpopulations and neighborhood income, it is found that there are, district-wide, significant correlations between test results and students' economic status, special education status, and English language proficiency. Furthermore, there is evidence that schools with a majority of students considered to be economically disadvantaged experience more pervasive testing failure. These findings contradict the premise of NCLB that we ought to ignore differences in student factors when evaluating instructional quality. The article suggests that while test scores may provide useful information regarding a given school, they are not valid for accountability purposes.

Keywords: high-stakes testing; teacher evaluation; school accountability; equity

Upon the publication of *A Nation at Risk* in 1983, there was a sense of urgency to address the many deficiencies of US public schools in the context of problems entrenched in our society. This seminal report spelled out the need, among other things, to emphasize the "twin goals of equity and high quality schooling" (The National Commission on Excellence in Education, 1983). In many ways, the direct, and appropriately bold, response on the part of lawmakers was the No Child Left Behind Act of 2001, that purported to ensure henceforth that all students would be taught effectively and that states, districts, schools, and teachers would be held accountable for the achievement of every student. In order to ensure this accountability, students would be assessed on standardized tests that would eliminate contextual considerations and ensure that we truly look at academic achievement.

According to the enforcement arm of NCLB—Adequate Yearly Progress (AYP)—schools are responsible for the achievement of all students as measured by yearly standardized tests. Based on their scores, students are or are not determined to be "proficient" in Reading and Math. Schools have to meet benchmarks that establish what percentage of their overall population ought to score Proficient. They also have to meet this benchmark for subpopulations of students—different ethnic groups, students who are economically disadvantaged, students with a disability (SWD), and English language learners (ELL), notably. Schools that do not make Adequate Yearly Progress (AYP) are publicized and required to take increasingly prescriptive corrective action. Through more recent initiatives such as Race to the Top, the Federal Government has also at least tacitly encouraged states to grade public school teachers according to their students' test scores, thus promoting the view that students' high stakes test results are the consequence of specific teachers' instruction.

In Washington, DC there has been a considerable push toward higher test scores, with an ongoing debate on how much teacher and administrator retention or salaries ought to be tied to test scores. In many ways, DC has been at the forefront of much of the push provided by NCLB and, consequently, also of the increasingly loud backlash against a law that many now claim actually contributes to the very inadequacies it was supposed to remedy.

Purpose

The purpose of this article is to examine how, in fact, AYP results—which are viewed as indicators of school and teacher quality—are a mirror of differences in the lives of children who are schooled in the same city. By comparing test scores across population groups, the study examines to what extent these scores can indeed—as the law mandates—serve as a way to hold schools accountable for educational quality, and to what extent they are instead a testimony for how not all students face the same challenges and a reminder that we need to consider students' differing situations in how we approach their schooling. To this effect, the article seeks to examine how pervasiveness of test failure is in fact related to pervasiveness in adversity beyond the classroom.

The study examines school proficiency rates on NCLB-mandated assessments for public middle school students in Washington, DC, in order to determine what non-instructional factors affect scores. The aim is to investigate whether it is indeed accurate to make judgments on a school's instructional success based on test scores—as NCLB and AYP measures currently dictate—or whether, instead, the reporting of test scores should be reconfigured to show what factors school staff in a given situation are in fact contending with.

The No Child Left Behind Act: Purpose and Central Provisions

When NCLB was enacted, equity issues in school quality had become a major public concern (Hall, Wiener, & Carey, 2003), with a sense that funding needed to be allocated to serve those schools most in need and, at the same time, that schools needed to be held accountable for what they achieved (McEntire, 2010). There was a sense that, especially in areas with large poor and minority populations, there was a general lack of rigor and focus on educational outcomes (McEntire, 2010; Hall et al., 2003). While many states were already putting in place standards to underlie curricula, few systematically, universally, and specifically assessed whether these standards were in fact being met (Linn, Baker, & Betebenner, 2002).

In many ways, NCLB was designed to be a catalyst for changes that would raise the general quality and international competitiveness of American education, promote equity, and hold schools accountable for reaching measurable outcome standards in ways common in the private sector (The Education Trust, 2004). In this way, the law was intended not to embody educational reform in itself, but to cause further reforms on the state, district, and school levels.

The impetus, or muscle, for getting schools to change their practice in this case was related to funding. Indeed, NCLB mandated that in order for states to receive funds allocated under Title 1, they commit themselves to getting all their students, regardless of demographic group, to be on grade-level by 2014. Accordingly, states became required to define benchmark goals to determine whether or not schools were making Adequate Yearly Progress (AYP) toward reaching this eventual universal proficiency. In this way, the AYP mandate set a standard for defining educational success that had previously been missing. The ambition of NCLB, then, was to force states to close the achievement gap and ensure that schools no longer had students who were in fact invisible—and that, indeed, every student counted.

Under NCLB, states have been held responsible for setting grade-level content standards and implementing an assessment tool that measures students' performance relative to these standards on an annual basis in grades 3 through 8 and at least twice in high school. A further requirement was established that specific scores be determined on the math and reading tests that mark whether students are or are not deemed to be proficient (The Education Trust, 2004). In recent years this notion of a single-point passing score that all AYP measures are based on has been challenged and a number of states are now working on using additional student improvement measures.

As required by NCLB, states set their own beginning proficiency target rate for 2003 based on previous baseline data and have been raising this target incrementally in order to reach 100% in 2014. A key provision of the law is that the same annual target rates are to be set for the total school population and for the different subgroups tested (by gender and ethnic group as well as low-income and limited English proficient students and students with disabilities). It has been required that at least 95% of students participate in the assessment. Finally, states were also asked to select one additional measure of academic progress, with most opting for attendance rates in elementary and middle schools and graduation rates for high schools.

Schools that do not meet these benchmarks but present a 10% decrease in students who did not score Proficient over the year before receive a 'Safe Harbor' (SH) designation that temporarily freezes their AYP status, indicating that while their progress is not deemed adequate, it is recognized that they are taking steps in the right direction. If a school does not make AYP for even one of its measures and subgroups, a series of increasingly prescriptive corrective measures are required, with a possible eventual restructuring of the school.

Measuring and Reporting AYP in the District of Columbia

The D.C. Comprehensive Assessment System (DCCAS), the high-stakes test that is used to assess at least 95% of students in grades 3 through 8 and 10 in Reading and Math, is given in all DC public schools and charter schools every spring. Using pre-determined single-score cut-off points, scores are determined as Below Basic, Basic, Proficient, and Advanced. Only

some students with severe disabilities are exempt from taking the test in its standard form and students with disabilities (SWD) and English language learners (ELLs) are permitted to have specific accommodations that do not significantly alter the test. Scores for ELLs who have been in the U.S. for less than one year are not counted toward AYP.

Schools are then issued a one-page "AYP Report" that is publicized in the local media. It indicates the percentages of students who were assessed and who scored Proficient or above. These are given for the total student population and for subgroups based on ethnicity, special education classification, ELL classification, and for students who are considered to be economically disadvantaged. Standard scores, analyses of the student population, and measures of statistical validity are not given. The AYP Report then concludes whether the school reached the statewide proficiency target for each student category. For 2011 these benchmarks were set at 73.69% Proficient in Reading and 70.14% in Math.

Assessing Student Differences Rather than Student Learning

A substantial body of research points to the impact that student background has on high-stakes test scores, thus in effect undermining the claim that these are indicators of instructional quality (Welner, 2005; Hershberg, 2008; Koretz, 2008; Jennings & Corcoran, 2009). According to Welner (2005), "the truth is that each—school and student—bears some responsibility, along with the state, the school district, the family, the community, peer groups, libraries, and various other people and institutions including the federal government. And the truth is that if it were possible to measure the actual contributions of each to student test scores, we would find varying proportions for each community, family, student, teacher, and school. A more rational NCLB would acknowledge both of these truths." What this means, then, is that AYP essentially fails to do precisely what it was designed to do, which is to provide a fool-proof tool with which to assess the results of instruction and hold schools accountable for student learning (Koretz, 2008; Martin, 2011).

Economically Disadvantaged Students, English Language Learners, and Students with Disabilities

Research indicates that some of the population groups that are specifically singled out in assessment of school performance are likely to receive lower scores across schools. Indeed, there is evidence that schools with high percentages of students who qualify for free and reduced lunch receive lower average scores than do schools with wealthier students (Escamilla, Mahon, Riley-Bernal, & Rutledge, 2003; Jennings & Corcoran, 2009). While there does not seem to be an obvious, single explanation, scores are found to be linked to socioeconomic factors and not just quality of instruction.

Inherent limitations in the validity of using test scores to measure educational quality are perhaps even more evident in the case of schools with large populations of ELLs. NCLB requires that after being enrolled in US schools for one calendar year all ELLs take the same English-language test as other students regardless of their English language proficiency. This has been a ground for concern as to the validity both of assessing ELLs in this way and in using the results to assess the performance of schools results (Government Accounting Office, 2006; Menken, 2008). It has been argued that for ELLs, in fact, any such standardized assessment of academic achievement is instead above all a test of English language proficiency (Menken, 2008, 2009). The consequence, then, is that schools are at a disadvantage relative to AYP if they even have a subpopulation of ELLs. There are indications that students with limited English proficiency have significantly lower scores than their peers (Government Accounting Office 2006; Abedi, 2009), which means that ultimately the AYP provision punishes schools for serving large numbers of ELLs (Escamilla et al., 2003; Ga'ndara & Baca, 2008).

Similarly, schools are hurt in terms of AYP for serving large numbers of students with disabilities (SWD). Abedi (2009) finds that special education students score significantly lower than their general education peers and Cole, Eckes, and Swando (2009) point out how common it is for schools not to make AYP only because of the test results of their SWD subgroup. Indeed, it has been found that standardized test results for special education students as a group are especially meaningless given huge gains and drops from one year to another that essentially make it impossible to trace a significant performance trend (Thurlow, Quenemoen, Altman, & Cuthbert, 2008). In fact, the very notion of standardized tests, based as they are on premises of homogeneity and reliability of testing behavior, seems to be at odds with the notion of disability. Because students are typically performing at least two years below grade level to even be labeled as having a disability, they would be required to make much faster progress than other students in order to make AYP, which seems especially nonsensical given the particular challenges they face (Eckes & Swando, 2009). It has also been shown that the assessment accommodations special education students typically receive have little impact on their performance (Bowen & Rude, 2006). As is the case for students considered to be economically disadvantaged and for ELLs, analyses of test results of SWD confirm that it is largely the nature of the student population rather than the quality of instruction that determine whether a school makes AYP (Eckes & Swando, 2009).

Testing for Teacher Accountability

Because of this close relationship between test scores and demographic factors, there have been calls to

abandon attempts to tie assessment results to teacher evaluation. In a discussion of value-added measurements, Au (2010) explains that while it appears that teacher quality has some effect on test scores, this cannot be accurately interpreted to mean that individual teachers are responsible for the results of individual students or that standardized tests are a simple reflection of instruction. In their study on using test score gains to evaluate teacher performance, Schochet and Chiang (2010) remark that over 90 percent of variation in scores is attributable to factors specific to the student and unrelated to the teacher. Baker et al. (2010) also point to yearly fluctuations in teachers' test results as evidence that students' test scores do not reliably reflect the quality of a specific teacher.

Research Questions and Methodology

A major emphasis of NCLB has been to hold schools accountable for all students' learning outcomes as measured on yearly standardized tests, eventually reconstituting schools that persistently fail to make AYP over time. A central question, of course, becomes whether this entails holding schools accountable for factors they have no control over. Caine (2011), writing about New York City schools, renders this logic as follows:

> Close a failing school with many failing kids and it's the end of the problem, right? New York City will have no more failing students, no more gangs, no more struggling families, no school-age children living in shelters, no more students with post-traumatic stress disorder, no more undiagnosed learning disabilities. (p. 50)

In order to examine the relationship between high stakes assessment results and contextual factors, available DCCAS results were collected for all DC public schools and charter schools that served grades six through eight (from the DC Office of the State Superintendant of Education, 2011). Three schools were omitted from the data because all or almost all of their students took an alternative assessment especially designed for students with severe disabilities instead of the DCCAS. Data were collected from a total of seventy schools, thirty-two belonging to the DC public school system and thirty-eight to the DC public charter school system. It was decided to focus specifically on the middle grades because not all elementary and high school grades are tested. However, not all grades six through eight in DC are served in separate middle schools. Instead, some are grouped with elementary or high school grades. Table 1 presents the breakdown of schools focused on by the grade range they served the year of the study. A limitation of the data, then, is that not all school test results reflect scores from exactly the same grades.

In order to understand whether contextual factors, as opposed to instructional quality, are significantly related to school-wide test scores, the following research questions are considered:

1) Is there a significant difference in test scores if students have what the AYP report singles out as a 'challenging factor'? It is to be noted here that NCLB does not mandate reports for groups of students considered not to have such 'challenging factors.' Instead, these are only given as embedded in the total school population.
2) Which supposed "challenging factors" have the most significant impact on school-wide scores?
3) Among schools that did not make AYP, there is a difference in pervasiveness of test failure based on whether they failed because of the scores of the total student population or only because of particular subgroups. How do such differences in pervasiveness of test failure correlate with the percentage of students considered to have "challenging factors"?
4) In the case of neighborhood schools, are there differences in test scores across neighborhoods with different mean family incomes?

For each school, proficiency scores were considered for the total population, students with economic disadvantages, SWD, and ELLs. Subtracting the number of each of these subpopulations from the total population of each school, the researcher also arrived at scores for students without economic disadvantages, students without disabilities, and students who were not ELL. Furthermore, schools were categorized according to their neighborhood cluster. Median family income for each neighborhood cluster was noted and schools were sorted accordingly.

Table 1
Number of Schools by Grades Served and Grades Tested

Grades served	PK-12	PK-8	K-8	3-8	4-8	5-8	6-8	6-9	7-8	6-12	7-12
Number of schools	1	36	1	1	3	4	16	2	4	1	1
Grades tested	All school levels	All elementary and middle school			Some elementary, all middle school		All middle school		Some middle school	All middle and high school	Some middle, all high school

Statistical analyses were then performed to test for significant relationships between each of these subpopulations and variables and the overall test results.

Data Presentation

Differences Among Subpopulations

Disaggregating and comparing the means of proficiency rates of students who were and were not found to have a particular "challenge," it is clear that students not categorized into one of the AYP-defined subgroups almost always have consistently higher proficiency rates, with a highly significant measure of statistical confidence (see Table 2). The one exception is Math scores for ELLs, which are not significantly different from scores of students who are not ELLs. Perhaps Math is a testing area where language proficiency matters less as a contextual factor. The limitation of these findings, however, is that while the public data allow us to calculate the percentages for students without specific "challenges" based on the subgroup data that are presented, the AYP reports do not allow us to determine results for students who do not have any "challenges." Those students, then, where obvious contextual factors might have the least impact on test scores, are those who are the most effectively hidden in the data. In order to highlight this omission, the table below features a row for these students that is of course left blank since the data are not made available to the public.

Most Significant Challenges

SWD are the subpopulation that is lowest in absolute terms and in relationship to their peers (15.7% Proficient contrasted with 40.12% in Reading and 21.96% Proficient contrasted with 52.72% in Math), suggesting the likelihood that the very fact that they have a disability—irrespective of particulars of instruction in a particular school—has a significant impact on their scores. As shown in Table 3, however, students with economic disadvantages constitute the majority of students in most schools (sixty-four out of seventy) and a large minority in almost all others (five out of six). In addition to scoring lower than students who are not considered economically disadvantaged within the same schools, they also score significantly lower if one simply compares the proportion of all students with economic disadvantages who score Proficient with the proportion of students who are not economically disadvantaged and who do so (with a Z value of 27.96 for Reading and 23.59 in Math, and a P value of less than 0.000 for both).

Table 2
Average School DCCAS Proficiency Rates by Student Subgroup (in schools where subgroups>25)

"Challenge"	Subject tested	Number of schools	Average subgroup Proficient score (students with "challenge" vs. student without)	Probability of no difference resulting from a one-tailed TTest (α=.05)
Economically disadvantaged vs. not economically disadvantaged	Reading	56	42.66% vs. 51.9%	P=0.003
	Math	56	46.85% vs. 53.89%	P=0.028
Students with disabilities vs. students without disabilities	Reading	46	15.7% vs. 40.12%	P<0.0001
	Math	46	21.96% vs. 52.72%	P<0.0001
English language learners vs. fluent English proficient students	Reading	14	38.52% vs. 57.74%	P<0.0001
	Math	14	47.49% vs. 58.69%	P=0.073 No significant difference
Students with an identified "challenge" vs. students without an identified "challenge"	Reading	Data not made public		
	Math			

Table 3
Prevalence of Students Considered Economically Disadvantaged

	Schools where they constitute a testing subgroup (>25 students)	Schools where they constitute the majority
Students with economic disadvantages	69	64
SWD	45	1
ELLs	14	0

Given their lower scores, the contextual challenges of students with economic disadvantages are likely to have a significant negative impact on total scores for schools as a whole. Simply put, the poverty of most of the students in most schools, as determined on the AYP reports, significantly hurts the testing profile of the entire school district.

Differences in Pervasiveness of Test Failure

While the law itself does not discriminate between reasons for not making AYP, there are arguably important practical differences between schools whose determination of failure on high stakes tests is pervasive, defined as school-wide, and others where low scores are limited to certain segments of the population. This has implications for the kind of improvement plan schools are held accountable for implementing following a failure to make AYP. If the school fails for the score of its total population, the whole program may need to be revised, whereas if it fails for only one specific subpopulation, specialized services might be targeted. There is also a difference in the public portrait that AYP paints of a school. If one has a child who does not have a disability, one may be reluctant to enroll her/him in a school where the total population is determined to have failed, yet one may have fewer misgivings if the school failed only because of the scores of its special education students.

Scores are therefore also examined to determine to what extent such relative pervasiveness of AYP failure is related to the makeup of the student population. Here, too, the connection is confirmed, although its significance is hard to ascertain given the small sample size. Simply stated, it may be that the less students in a school are subject to recognized contextual challenges, the less pervasive AYP failure is. The more challenges students enter the school with, the more likely it may be that the school will completely fail in terms of AYP (see Table 4). Data here are analyzed in regard to the percentage of students in a school who are considered economically disadvantaged. Because of the smaller number of schools that have SWD and ELL sub-groups, this analysis is not performed to account for those factors.

As a rule, schools where less than half the population is considered economically disadvantaged meet the AYP target for the total population. In cases where they do not make AYP it was mostly because of specific subgroups only. Conversely, the large majority of schools where most students are labeled economically disadvantaged fail to make AYP for their total population. AYP failure, then, is often central and pervasive for schools where most students are poor and peripheral, limited to small subgroups for schools where students are wealthier. From an outside perspective, then, pervasiveness of testing failure mirrors the economic context of the students.

Differences Across Neighborhoods

Enrollment in the schools belonging to the local school system is primarily based on the neighborhood of residence. To account for income segregation among neighborhoods, the study examines the relationship between the mean income of the school's neighborhood and test scores. For this the researchers used data on average family income for the neighborhood cluster of each school (Urban Institute, 2011). Overall, these figures range from $41.510 to $288,541. Performing a regression analysis to examine the relationship between test scores and average family income for the neighborhood cluster, one finds a very significant correlation (with an F-statistic of less than 0.0001 for both Reading and Math). In other words, the neighborhood of the school and the income of its population are significant factors in determining the proficiency scores of the school.

Table 4
Pervasiveness of AYP Failure as it Relates to Economic Disadvantage

Percentage of students considered economically disadvantaged	0-25%	25.1-50%	50.1-75%	75.1-100%	Test Statistic (∞=.05)
Schools that did not make AYP for total population in both Reading and Math (40 out of 70 total)	0	1	12	27	Chi^2=47.4 P<0.0001
Schools that did not make AYP for total population in either Reading or Math (19 out of 70 total)	0	0	6	13	Chi^2=24.16 P<0.0001
Schools that did not make AYP only for subpopulations in Reading and/or Math (13 out of 70 total)	2	5	3	3	Chi^2=1.46 P=0.6915. In addition to confirming the null hypothesis, the results are not reliable due to the value in each frequency.

Discussion and Implications

There are legitimate questions as to exactly what information assessment scores give us. Based on this study, and on others, it seems that some of this information has to do with the impact of non-instructional factors on student performance on specific academic tasks. If we are to better understand how we ought to teach and understand the needs of our students, this information might be valuable indeed—as formative information to help us consider our practice and differentiate for specific groups of students in specific contexts. However, given that this information pertains to student situations, it does not qualify as an appropriate tool for assessing the performance of their teachers. Unless we want to hold teachers at least partly accountable for their students' family income, home language, and disabilities, it seems to be a mistake to use high-stakes test scores for teacher and school accountability purposes. If anything, low scores would seem more likely to drive schools to try to change their students than their instructional practices—or, in the case of teachers, to change schools.

Indeed, these assessment scores tend to describe that economic segregation, for example, is alive and well and that it is embodied in our public schools. Given a choice, then, AYP reports as they are formulated would seem to encourage families and teachers to try to move out of poorer neighborhoods and into wealthier ones. This cannot be what *A Nation at Risk* had in mind and, one would hope, not what No Child Left Behind intended, either.

Perhaps this misuse of test scores exemplifies what could be a general tendency to gloss over contextual factors that underlie our inequities rather than address them. Indeed, within each of these AYP "subgroups" there are numerous dynamics that need to be addressed and that belie the notion that everyone can simply be tested in the same way in the hope that the tests always mean the same thing.

This reductionism to single-method quantification is a way to deny context—at best out of the mistaken belief that considering contextual factors would keep us from getting to the core of the issue. Couldn't it be, however, that these contextual factors are in fact the core of the issue that needs to be described, highlighted, discussed, and addressed—that they are a part of the formative information teachers ought to consider as they plan instruction? And isn't it also true that by focusing primarily on judging schools not for what they do but for what they contend with constitutes yet another instance of a lack of support to institutions we say we think are important? Dunn (2005) expresses it with the following:

> The stresses that students and teachers encounter in schools today should evoke compassion and admiration from the public; unfortunately, quite the opposite occurs, and this troubles me even more. Test results are released and inner-city students and their teachers are ridiculed in bold headlines. (p. 179)

This study therefore suggests the following recommendations:

- These differences need to be understood, acknowledged, and highlighted, rather than glossed over for the sake of accountability if NCLB is to truly perform its self-declared duty of ridding our public schools of their inequities.
- Going along with that, the social inequities that are reflected in the test results should not be ignored in order to focus on the work of educators. Instead, they should be regarded in order to reemphasize that if we truly want educational success for all, inequities outside of the world of school need to be addressed head-on.
- High-stakes tests should be considered as formative rather than summative assessment information to benefit instruction.
- If we want to examine the work of teachers and the impact of instruction, standardized tests are the wrong measure.
- By highlighting rather than hiding contextual factors, we may actually use this kind of data to identify how a school is unique and define goals, expectations, and practices for that particular context. Instead of forcing us into a one-size-fits-all model, high-stakes assessment scores can help us differentiate for individual schools and children.
- Non-"challenged" students ought to be counted as a sub-group.
- Schools should cease to be judged on the demographic they teach.
- Effectively, assessments as used and put in practice promote the very inequities NCLB was designed to redress.
- For further research, the possible connection between pervasiveness of test failure and demographic makeup of the student population ought to be examined further on a larger scale.

The point is not that the testing data are necessarily meaningless or shouldn't be used as one of many tools to inform our educational planning. They describe something, some kind of a reality that may be pertinent. What exactly this reality is certainly warrants further investigation and discussion. The point is that they are useful as one descriptor among many, as a valuable albeit limited and flawed informant. They are certainly not, however, meaningful as a tool for evaluation. We need to move away from our frenzy to evaluate and try a little more to describe and understand the reality. On the contrary, it may be that this emphasis

on accountability actually discourages us from developing what Senge, Scharmer, Jaworski, and Flowers (2004) term "the courage to see freshly" because we are too worried about how this seen reality will reflect on us (p. 35). For this type of understanding, contextual factors in students' lives are crucial. For teacher evaluation, however, they are in the way, making us only want to dismiss them as a potential smoke-screen.

References

Abedi, J. (2009). English language learners with disabilities: Classification, assessment, and accommodation issue. *Journal of Applied Testing Technology, 10*(2).

Au, W. (2010). Neither fair nor accurate. *Rethinking Schools, 25(2)*, 35-38.

Baker, E. L, Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers. (Issue brief, 278).* Washington, DC: Economic Policy Institute. Retrieved from: www.epi.org/page/-/pdf/bp278.pdf

Bowen, S. K., & Rude, H. A. (2006). Assessment and students with disabilities: Issues and challenges with educational reform. *Rural Special Education Quarterly, 25*(3), 24-30.

Caine, W, (2011). My failing school. *Rethinking Schools, 25(4)*, 48-51.

Cole, C. (2006). *Closing the achievement gap series: Part 3. What is the impact of NCLB on the inclusion of students with disabilities?* Bloomington, IN: Center for Evaluation and Education Policy, Indiana University.

D.C. Office of the State Superintendant of Education (2011). *AYP reports*. Retrieved from: http://www.nclb.osse.dc.gov/index.asp

Dunn, B. (2005). Confessions of an underperforming teacher. In Nieto, S. (Ed.), *Why we teach*. New York, NY: Teachers College Press.

Eckes, S. E., & Swando, J. (2009). Special Education subgroups under NCLB: Issues to consider. *Teachers College Record 11*(11), 2479-2504.

Escamilla, K., Mahon, E., Riley-Bernal, H., & Rutledge, D. (2003). Stakes testing, Latinos, and English language learners: Lessons from Colorado. *Bilingual Research Journal, 27(1)*, 26-50.

Ga'ndara, P., & Baca, G. (2008). NCLB and California's English language learners: The perfect storm. *Language Policy, 7,* 201-216.

Government Accounting Office. (2006). *No child left behind act: Assistance from education could help states better measure progress of students with limited English proficiency*. Washington, DC: Author.

Hall, D., Wiener, R., & Carey, K. (2003). *What new "AYP" information tells us about schools, states, and public education.* Washington, DC: The Education Trust.

Hershberg, T. (2005). Value-added assessment and systemic reform: A Response to the challenge of human capital development. *Phi Delta Kappan*, *87(4)*.

Jennings, L. J., & Corcoran, S. P. (2009). "Beware of geeks bearing formulas": Reflections on growth models for school accountability. *Phi Delta* Kappan*, 90(9),* 635-639.

Koretz, D. (2008). A measured approach. *American Educator, Fall 2008*.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31(6),* 3-16.

Martin, P. C. (2011). Selecting one story and hiding others: How AYP chooses the portrayal of a school. *Current Issues in Education*, 14(1). Retrieved from http://cie.asu.edu/

McEntire, N. (2010). Legacy of No Child Left Behind. *Childhood Education, 87(1),*57-58.

Menken, K. (2008). *English learners left behind: Standardized testing as language policy.* Clevedon, Avon: Multilingual Matters.

Menken, K. (2009). Policy failures: No Child Left Behind and English language learners. In S. Groenke & A. Hatch (Eds.), *Critical pedagogy and teacher education in the neoliberal era: Small openings*. Berlin, Germany: Springer.

Schochet, P. Z., & Hanley, S. C. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE, 2010-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Senge, P., Scharmer, C. O., Jaworski, J., & Flowerd, B. S. (2004). *Presence: Exploring profound change in people, organizations, and society*. New York, NY: Currency Doubleday.

The Education Trust (2004). *The ABCs of "AYP." Raising achievement for all students.* Washington, DC.

Thurlow, M., Quenemoen, R., Altman, J., & Cuthbert, M. (2008). *Trends in the participation and performance of students with disabilities* (Technical Report, 50). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

The National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform. A report to the nation and*

the Secretary of Education. United States Department of Education. Washington, DC.

Urban Institute (2011). *Neighborhood info DC.* Retrieved from: http://www.neighborhoodinfodc.org/nclusters/nclusters.html

Welner, K. G. (2005). Can irrational become unconstitutional? NCLB's 100% presuppositions. *Equity & Excellence in Education*, *38*, 171–179.

Article Citation
Martin, P. C. (2012). Misuse of high-stakes test scores for evaluative purposes: Neglecting the reality of schools and students. *Current Issues in Education*, *15*(3). Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view/1061

Author Notes
Peter Clyde Martin, Ed.D
Ithaca College
Phillips Hall 194C
953 Danby Road
Ithaca, NY 14850
(607) 274-1076
pmartin@ithaca.edu

Dr. Martin is on the faculty of the Education Department at Ithaca College. He holds a doctoral degree in Bilingual Special Education from The George Washington University. His research interests focus on the areas of teacher education and differentiated instruction, teacher collaboration, educational equity, vision-based and transformational schooling, and serving the needs of English language learners considered to be at risk of educational failure.

Current Issues in Education
Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Manuscript received: 8/26/2012
Revisions received: 11/6/2012
Accepted: 11/8/2012

Current Issues in Education

Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Volume 15, Number 3 December 1, 2012 ISSN 1099-839X