20 NCPEA 11

BLAZING NEW TRAILS:
Preparing Leaders to Improve Access and Equity in Today's Schools

THE 2011 YEARBOOK OF THE NATIONAL COUNCIL OF PROFESSORS OF EDUCATIONAL ADMINISTRATION

BETTY J. ALFORD, ***Editor***
GEORGE PERREAULT, ***Associate Editor***
LUANA ZELLNER, ***Associate Editor***
JULIA W. BALLENGER, ***Assistant Editor***

DES*tech* Publications, Inc.

PRO>ACTIVE PUBLICATIONS

Lancaster, Pennsylvania

Structured Inequity: The Intersection of Socioeconomic Status and the Standard Error of Measurement of State Mandated High School Test Results

Christopher H. Tienken

Assessment-driven education policies are in place in all 50 states in America. The reauthorization of the 1965 Elementary and Secondary Education Act (ESEA, P.L. 89-10), known as the No Child Left Behind Act (No Child Left Behind [NCLB PL 107-110], 2002), signed into law on January 8, 2002, cemented test-based policy making into the education landscape during the first decade of the new millennium. The introduction of the Race To the Top (RTTT) competitive grant program administered by the United States Department of Education (USDOE), and the report, *A Blueprint for Reform, The Emergency and Secondary Education Act* (ESEA Blueprint) *Reauthorization* (United States Department of Education, 2010), combined with stated support for the Common Core State Standards by 49 states and territories added more pressure to continue the policy practice of using standardized test results as the sole or deciding factor to evaluate student achievement and public education effectiveness.

The policy and practice of using results from statewide standardized tests to evaluate students and education quality is not new (Education Commission of the States, 2008). Georgia, Texas, Florida, and Louisiana and cities, such as New York and Chicago have used results from standardized state tests to make grade promotion decisions about students for some time, and 23 states used statewide exams to determine high school graduation eligibility in 2009. The practice of using high school exit exams as the deciding factor on whether a student can receive a standard diploma began over 30 years ago in 1978. By 2012, Arkansas, Maryland, Pennsylvania, and Oklahoma might also use exit exams, bringing the total to 27 states (Education Commission of the States, 2008).

School administrators in the 50 states are encouraged to make data-driven decisions based on the results of state mandated tests (Booher-Jennings, 2005; Leithwood, Louis, Anderson, & Wahlstrom, 2004; Weiss, 1998). For example, the word "data" appears 230 times in the NCLB Act legislation. The word data appears 16 times, almost once every-other page, in the report (ESEA Blueprint) (United States Department of Education, 2010), and the RTTT program requires administrators to use results from state mandated tests to make decisions about student achievement and teacher effectiveness. Every state education agency has at least one statement related to data-driven decision making on its official web pages, and most have special pages related to data reporting from statewide tests of academic skills and knowledge. School administrators use state assessment results as data to make decisions and judgments about such things as teacher effectiveness, student achievement, and program effectiveness (Burch, 2005; Heubert & Hauser, 1999; Penfield, 2010; Roderick & Engel, 2001; Tienken, 2008).

All results from statewide tests of academic skills and knowledge contain technical flaws that should preclude them from being used as the only data point or as the deciding factor

Christopher H. Tienken, Seton Hall University

Author Note: The author would like to thank Minmin Fan and Olivia Rodriguez for their assistance with this manuscript.

to make high-stakes decisions about individual students, such as for high school graduation or grade promotion (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999; Joint Committee on Testing Practices [JCTP], 2004); yet, the practice continues. Unintended social and education consequences of using the results from one state mandated high school exam to make important decisions can include students being retained in grade (which increases the chances of not completing high school), placement in low-level course sequences (which increases the chances of not completing high school), having to take the test again and endure a semester or year of a test preparation course, mandating students to go through an alternative assessment procedure, not receiving a standard high school diploma, or being denied graduation.

Not graduating from high school or being denied a high school diploma can trigger a series of negative events in terms of life-long consequences. As a group, adults who do not hold a high school diploma earn between $7,000 to $10,000 less per year than adults who have a high school diploma (Cheeseman Day & Newburger, 2002). Individual earnings can be related to a person's long-term health with the difference in life expectancy between middle class and wealthy Americans is almost five years more than for poor Americans (Thomas, 2010). Depressed earnings result in lower tax receipts, and they are also associated with higher public medical costs, greater rates of incarceration, and greater use of the welfare system (Levin, 2009). Negative consequences are associated with use of the test results to make potentially life-altering decisions about students (Messick, 1995, 1996).

Test Score Validity and Misinterpreting Results

Messick (1995, 1996) cautioned psychometricians that the traditional view of validity as three distinct categories, construct, content, and criterion, is ill-suited to explain the potential negative social and education consequences of test-score misinterpretation. He proposed a more comprehensive and progressive view of validity that integrated criteria and content validity with intended and potential unintended consequences associated with high stakes testing within the construct validity framework. Messick (1995) placed the intended and unintended social and education consequences of test score interpretation or score misuse as an aspect of construct validity and not as its own category of validity. Messick's proposal suggested that those who create and use high stakes tests should weigh the possible intended and unintended consequences to children before enacting a testing program. The integrated view of construct validity allows school administrators and policymakers to consider social and education consequences in the validity discussion and potentially make more informed policy decisions.

One troubling technical characteristic associated with construct validity and the use of the results from state mandated high school tests to make potentially life-altering decisions about individual students is conditional standard error of measurement (*CSEM*) and its effect on individual test-score interpretation. The reported results of individual students might not be the actual or true scores. The *CSEM* is an estimate of the amount of error the user of test results must consider when interpreting a score at a specific cut-point or proficiency level or when making a high-stakes decision based on the test score (Harville, 1991). Think of *CSEM* as the margin of error reported in political polls (e.g. + or – 7 points): The individual student-level results from every large-scale state standardized test have a margin of error. The *CSEM* describes how large the margin of error is at the various proficiency cut-points and how much the reported test results might differ from a student's true score.

For example, if a student receives a reported scale score of 546, and there are + or – 12 scale-score points of *CSEM at the proficiency cut-point*, then the true score could be located somewhere within the range of 534 to 558, and the student could be expected to score within that range if he or she took that test again. If that state's proficiency cut-score is 547, then the student is rated *not proficient* based on his or her reported score if the State Education Agency personnel (SEA) do not account for *CSEM* in some way in the proficiency calculations, even though the student scored within the error band, only one point away from proficiency. This is especially troubling when the single test score determines if a student can graduate high school or receive a standard diploma, as it does in 23 states (Education Commission of the States, 2008).

Problem

A more focused problem appears at the confluence of *CSEM,* score interpretation policy for high school exams in the 50 states, and the documented effects of group membership in the Economically Disadvantaged (ED) subgroup on ultimate student achievement. Students eligible for free or reduced lunch, known in many states as *Economically Disadvantaged*, score as a group statistically and practically significantly lower on statewide high school exams, and state exams in all other grade levels, than their peers who are *Non-economically Disadvantaged* (Non-ED). Students in the ED subgroup are more likely, as a group, to be affected negatively by misinterpretations of score results due to *CSEM* that cause them to be labeled as not proficient because they score closer to their state's proficiency cut-score. There has been little empirical research published since the inception of NCLB that describes the amount of error present in high school state standardized test scores for language arts and mathematics. Even less literature exists that attempts to account for the number of students potentially harmed by SEA policies that do not account for the error inherent in the individual scores of students.

The purpose of this chapter is to (a) describe the practical significance of the differences in results on state mandated high school exams in language arts and mathematics between students categorized as *Economically Disadvantaged* (ED) and those not categorized as ED, (b) determine the number of students potentially miscategorized as not proficient due to *CSEM,* and (c) describe the policy options available to state education agency personnel and school leaders.

Research Questions and Significance

Three questions guided the study: (a) How do SEA personnel attempt to remedy the imprecision issues posed by *CSEM* on the interpretation of reported individual student test scores?; (b) What is the practical significance (effect size) between high school exam results on the language arts (LA) and mathematics (M) sections for students designated as economically disadvantaged (ED) and those not ED?; and (c) Approximately how many students are potentially mislabeled as less than proficient on state LA and M exams due to *CSEM*? The results of this study provide leverage, on which to advocate for policy adjustments. Education policy and high-stakes testing schemes continue to take shape at the federal level, and the informed discussion of *CSEM* should be a priority topic.

RESEARCH AND LITERATURE ON HIGH SCHOOL EXAMS AND THEIR RESULTS

This section provides an overview of the characteristics of the literature on the topic of state mandated high school standardized tests and CSEM. I conducted an initial Internet search and used Boolean techniques to explore the literature on the topic of state mandated high school standardized tests and *CSEM*. The search terms included *conditional error of measurement* and *state mandated tests*, *measurement error* and *high school exam*, and *high school state exams* and *conditional standard error of measurement*. The initial search produced three types of results: (a) non-empirical literature, (b) empirical literature, and (c) psychometric technical documents and professional standards for testing. A second search was conducted using the AERA and Education Policy Analysis Archives journal databases. The results of the second search also produced results that fit into the three categories above.

Non-empirical Literature

The non-empirical literature ranged from advocacy, policy briefs, and editorials published by think-tanks and researchers who support the practice of using state mandated high school test results to make high-stakes decisions about children (e.g., Achieve Inc., 2008; Education Commission of States, 2008; Freedman, 2004; Greene & Winters, 2004; Hanushek & Welch, 2006) to literature of those who opposed the practice (e.g., Fairtest.org, 2008; Neill, 1997; Ohanian, 2001). Although the non-empirical literature might not rise to the level of empirical research as defined by Haller and Kleine (2001), it has influenced education policy in the past (e.g., Goals 2000, NCLB, Achieve, Inc. and its *American Diploma Project, Common Core State Standards, RTTT, ESEA Blueprint for Reauthorization*). There is little discussion about the *CSEM* in the non-empirical literature.

Empirical Literature

In a related study (Tienken, 2009), a review of empirical literature on *CSEM* issues and high school exams revealed 53 peer-reviewed articles with the terms "high school exam." A Boolean search with the terms *conditional standard error of measurement* and *high school exam* did not result in a peer-reviewed article that reported the actual scale-score *CSEM* present in high school exams used nationally or reported directly on the influence of *CSEM* on interpretation of the results. However, three contradictory claims about the influence of high-stakes high school exams on student achievement and graduation rates surfaced. For example, in terms of high school exit exams (in use in 23 states), the literature suggested they (a) improve overall achievement and graduation rates (Stringfield & Yakamowski-Srebnick, 2005); (b) suppress overall achievement and graduation rates, and have negative unintended consequences, especially for minorities (Hursh, 2007; Lee & Wong, 2004; Vasquez Heilig & Darling-Hammond, 2008); or (c) provide mixed, uneven, or inconsistent results (Allensworth, 2005; Clarke, Shore, Rhoades, Abrams, Miao, & Li, 2006).

Standards for Education Testing

Authors of *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and the *Code of Fair Testing Practices in Education* (JCTP, 2004) present specific standards and recommendations for test developers, test takers, and those who use

test results to make decisions about children. The standards and recommendations cover test construction, fairness in testing practices, appropriate documentation of technical characteristics of tests, and other related topics. Both publications make specific recommendations for how state personnel and school leaders can address *CSEM* in the context of high-stakes testing. I chose to focus on the *Standards* instead of the *Code* because the three largest organizations (in terms of membership) associated with testing produced the *Standards* (APA, AERA, and NCME, 1999). They provide specific guidance for developers and users of high stakes testing programs, and the working group who produced the *Code* included members of the three *Standards* organizations, and many recommendations contained in the *Code* are included in the *Standards*.

Specific statements related to construct validity, as defined by Messick (1995, 1996), and measurement error are listed in Part I and Part III of the *Standards* (AERA, APA, & NCME, 1999). The authors of the *Standards* concurred with Messick (1995; 1996) when they wrote:

> Measurement error reduces the usefulness of measures. It limits the extent to which test results can be generalized beyond the particulars of a specific application of the measurement process. Therefore, it reduces the confidence that can be placed in any single measurement. (p. 27)

The authors recommended that error and its sources be reported, stating, "The critical information on reliability includes the identification of the major sources of error, summary statistics bearing on the size of such error…." (p. 27). The authors of the *Standards* explained why test developers and users (i.e., SEA, school administrators, policy makers) must report and be aware of the *CSEM* at the proficiency cut-score levels on tests:

> Mismeasurement of examinees whose true scores are close to the cut score is a more serious concern. The techniques used to quantify reliability should recognize these circumstances. This can be done by reporting the conditional standard error in the vicinity of the critical value. (p. 30)

Table I includes the applicable macro-standards, statements, and paraphrased recommendations related to error and reporting. Authors of the *Standards* provide overall guidance on interpretation and score precision stating, "The higher the stakes… the more important it is that the test-based inferences are supported with strong evidence of technical quality" (p. 139).

Theoretical Perspectives for Using Statewide High School Exams as High-Stakes Indicators of Achievement

Advocates of high school exams generate policy frameworks and proposals from the rationalistic and behaviorist fields of education psychology. The proposals are operationalized via state education policies that use positive reinforcement and punishment, also known as carrots and sticks. Bryk and Hermanson (1993) termed this an *instrumental use* model. Norris, Leighton, and Phillips (2007) termed it the *Stakes Competency Model*. The theory is that a policy body develops a set of expected education outcome measures (e.g. state standards) and monitors the relationship between the measures and school processes through a high stakes statewide standardized test, and then attempts to change behavior of those in the system through external force. The standardized testing measures rest upon arbitrarily set proficiency

bands and external control (e.g., threats of state takeover, vending the school to an education management corporation, or state monitoring).

Table 1. Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) Related to Test-Score Precision and Conditional Standard Error of Measurement.

Standard	Standard Statement	Recommendations
2.2	"The standard error of measurement, both overall and conditional…, should be reported…in units of each derived score" (p. 31).	The CSEM is important in high school exit exam situations due to the consequence of imprecision.
5.10	"…those responsible for the testing programs should provide appropriate interpretations. (They) should describe …the precision of the scores, common misinterpretations of test scores…" (p. 65).	Score precision should be illustrated by error bands or potential score ranges for individual students and should show the CSEM.
6.5	"…When relevant for test interpretation, test documents ordinarily should include item level information, cut scores…the SEM…" (p.69).	The SEM should be reported.
7.9	"When tests or assessments are proposed for use as instruments of social, educational or public policy, …users …should fully and accurately inform policy-makers of the characteristics of the tests…" (p. 83).	Precision is an important issue… Users should report the amount of error present in scores.

Advocates of high-stakes testing policies postulate that high-stakes exams cause students and teachers to work harder and achieve more because the tests create teaching and learning targets that have perceived meanings to both groups. There are underlying assumptions that teachers and students do not already work hard and that one test can measure and provide information that is meaningful in terms of student achievement and systemic efficacy. Another example of the theory in policy includes the threats from State Education Agency's (SEA) to withhold funding for poor performance to compel school personnel to work harder because they do not want to lose funding. A similar version is the use of public castigation via the press and ratings and/or rankings of districts by SEA personnel to spur educators to work harder to achieve outcomes. This type of policy making philosophy is in line with Rational Choice Theory. But those who rely on Rational Choice Theory seem not to understand Reactance Theory: You push me, I push back, resist, and/or subvert.

Conversely, high-stakes exam opponents derive theoretical guidance from an enlightenment model based on self-determination theory (Laitsch, 2006). Creators of an assessment system based on an enlightenment model seek to foster greater discussion, study, and reflection of education practices based on the indicators of the assessment system. Standardized tests still play a part, but their uses and interpretations are different compared to those within an instrumental use model, and they are not high stakes in nature. The system includes multiple data points, both quantitative and qualitative. Greater use might be placed upon teacher grades or student grade point average, which have been shown to be a better predictor of first year college success than the SAT (Zwick, 2004).

DESIGN AND METHODOLOGY

I used a non-experimental, exploratory, descriptive cross-sectional design (Johnson, 2001) to answer the research questions. Data were collected between 2008 and 2010 from publically available state test technical manuals and databases. First, an Internet search of

SEA websites was conducted for the mathematics (M) and language arts (LA) exam technical manuals of the 50 states that use high school exams. I used the "search" function on each SEA site to locate the technical manuals and used Boolean search techniques and appropriate descriptors to find high school exam technical manuals. Formal emails were sent to the SEA testing coordinators to request the technical manuals if the manual was not posted on the SEA website. A second email was sent after two weeks if a reply was not received. In some cases, I called the assessment directors to ask for information. Technical manuals are supposed to be in the public domain as recommended by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999).

Some states posted multiple years and grade levels of technical manuals for each subject. I chose the most recent manual at the highest grade level if there were exams for multiple high-school grades. The manuals' most recent publication dates ranged from 2005 to 2009. If a state included Algebra I and Algebra II exams, the Algebra II exam was chosen because of the assumption that the Algebra II exam would represent more closely the higher level of high school math attainment. *SEM* and *CSEM* values for each test in each state where data were available were determined from a previous study (Tienken, 2009) on that subject, and they are listed in Appendix A. of this chapter.

Then, I searched each SEA website for information regarding cut-score setting methods and the treatment of *CSEM*. In cases in which information was incomplete or not included, the director of state testing was emailed to request the information. As per the *Standards for Educational and Psychological Testing* that type of information is supposed to be reported in the public domain. Finally, I used the publically accessible *Center on Education Policy* database that reports state test results for various subgroups of students across the 50 states. I created a table (see Appendix) to show the LA and M scores from high school exams for students in the ED subgroup and those in the non-ED subgroup. Glass's Delta formula was used to calculate the effect size difference (practical significance) in mean scores for the two groups.

It should be noted that the term Economically Disadvantaged is the term used in this chapter to describe those students designated by their states as eligible for either free or reduced priced lunches at school; this term was used most often in the literature and data reviewed. I am well aware of the potential weaknesses of relying on free or reduced lunch status as the primary indicator of a student's complete economic status (Harwell & LeBeau, 2010). Free/reduced lunch status is a blunt indicator of socio-economic status. There are meaningful differences between being eligible for free lunch as opposed to reduced lunch and those differences have varying influences on student achievement. Data from the National Assessment of Education Progress (NAEP) for M and LA results for Grades 4 and 8 suggest that students eligible for free lunch scored statistically significantly ($p < .05$) lower than students not eligible for free or reduced lunch. Conversely, there was not a statistically significant difference in scores between students eligible for reduced lunch and those not eligible for reduced or free lunch.

The free lunch category captures some of the effects of poverty whereas the reduced lunch category does not. However, states do not often separate achievement into the two distinct categories, and instead, report achievement as one category: free/reduced lunch. This designation masks some of the negative influences of poverty because the scores for students eligible for free lunch would be even lower than those in the category known as free/reduced lunch. The combined free/reduced lunch category does not allow for deep exploration of the effects of poverty because it includes students whose family income is up to $39,220, almost two times the federal poverty level income threshold.

The federal guidelines for determining eligibility contribute to the blurriness of the indicator. The guidelines have not been substantially updated since they were created in 1960s, and they do not take into account other factors that depress after-tax income that were never considered when the guidelines were created. Those factors include such things as costs for child care, health insurance premiums and related costs, variations in costs of living throughout geographic regions, transportation costs, and the influences and effects of living in an impoverished neighborhood (Harwell & LeBeau, 2010). Because more fine grained census data or definitions of economic status are not provided by states, I chose to use the data that were reported most widely.

RESULTS

Appendix A presents (a) the name of each state; (b) the most recently reported or estimated *CSEM* at the proficiency cut-point for the LA and M portions of the high school exams; (c) mean scale scores, standard deviations, and population sizes for students in the ED subgroup and those in the non-ED subgroup; (d) number of students potentially affected negatively (miscategorized as not proficient) due to not accounting for *CSEM* in the individual scores; and (e) effect size differences between the mean scale scores for the ED subgroup and Non-ED subgroup.

The range of *CSEM* at the proficiency cut-point for LA was 3.3 scale-score points to 89 scale-score points and the range of *CSEM* at the proficiency cut-point for M was 3.3 scale score points to 88 scale-score points. I am less concerned with the size of the error because each state uses a *hard and fast* cut-score. That means there is no accommodation for *CSEM*, almost as if it does not exist. If a state's proficiency cut-score is 200, as it is in New Jersey, and a student scores a 198, then that student is categorized as not proficient, even though there are approximately nine points of error at the cut-point on the New Jersey tests. Therefore, even one point of *CSEM* can cause misinterpretation and miscategorization of student performance because SEA personnel do not account for *CSEM* in individual test results.

Every SEA provided at least two opportunities for students to take and pass the high school exam. The mode was three testing opportunities. That was the SEA-preferred mechanism to deal with not accounting for *CSEM* in the individual student scores. None of the SEA reporting policies awarded the CSEM to the student. Only two states (4%) stated that they attempted to account for the *CSEM* in the score setting process, but further review of their processes, as stated in their technical manuals, revealed inconclusive methods. One state reported that the CSEM was accounted for by setting the initial proficiency cut score lower to account for the error. That just moves the problem to a different cut-score. A more appropriate method would be to award the students the scale score *CSEM*, the margin of error if you will, at the proficiency cut-score to their results. None of the SEAs account for the *CSEM* by awarding the student the theoretical higher score, the score at the top-end of the error band.

The Intersection of Not Addressing *CSEM*, Being Economically Disadvantaged, and Structured Inequity

More than one quarter, 13/50, (26%) of the SEA did not report mean scale scores for the ED and non-ED subgroups. For all states that did report those data, 37/50 (74%), there was no instance when the ED subgroup achieved a higher mean score on the LA or M portions of the high school state tests than the non-ED subgroup. In 37 states that reported data, the children in the ED subgroup scored closer, and in some cases, below the proficiency

cut-score for their respective states. In 12/37 states that reported data, children in the ED subgroup scored below their state's proficiency cut-score in mathematics. In 8/12 of those states, they scored within the *CSEM* band from proficiency, meaning that, as a group, the ED students in those 8 states (in 75% of the states where this occurred) would have achieved a mean score above the proficiency cut-point had the SEA personnel in those states accommodated for the *CSEM* in the individual student results. Instead, their mean group score fell below the proficient level making them candidates to be more likely miscategorized as not proficient, due to measurement error than their non-disadvantaged peers.

The data suggested that students in the ED subgroup are more likely to be affected by not accommodating for *CSEM* in test scores than their non-ED peers, and they are more frequently categorized as not proficient than if the error were addressed as recommended in the *Standards for Educational and Psychological Testing*. The students in the ED subgroup scored lower in LA and M in every state that reported data. They scored closer to their states' proficiency cut point in every state that reported data. They actually scored below their states' proficiency cut points in LA in 11/37 (30%) of the states that reported data and below in M in 12 states. Not accounting for error places the students in the ED subgroup at greater risk of being categorized mistakenly as not proficient.

By comparison, in only five states did the non-ED student subgroup score below their state's math proficiency cut-score. The non-ED student sub-group scored within the *CSEM* range in four of the five states. None of the non-ED students in any state scored below their state's proficiency cut-score in LA, whereas the children in the ED subgroup in 11 states scored below their states' LA proficiency cut-score and in 9/11 (82%) of those states, the students scored within the *CSEM* range on the LA test.

The achievement differences were striking in terms of scale scores and effect sizes. The effect size differences in mean achievement between the students in the ED subgroup and their non-ED peers ranged from 0.39 to 1.05 in LA and 0.36 to 1.02 in M. The effect size was 0.50 or higher favoring the non-ED in LA and M in 27/37 (73%) states that reported data. To put that into perspective, an effect size of 0.50 favoring the non-ED subgroup would be the difference between a student scoring at approximately the 67th percentile on a nationally norm-referenced test compared to a student scoring at the 50th percentile.

Number of Students Affected

I was able to locate or estimate the number of students in 23/50 (46%) students potentially affected negatively by not accommodating *CSEM* (i.e., being miscategorized as something less than proficient). An estimated 166,305 students were miscategorized at least once in an academic year as *less than proficient* on their statewide mandated LA test because of *CSEM* and the fact that SEA personnel do not account for it at the student level. Similarly, an estimated 164,982 students were categorized as *less than proficient* on their statewide mandated math test. It is unclear how many students who were miscategorized in M were also miscategorized in LA or vice versa. Because students in the ED group scored closer to their states' proficiency cut points more frequently than their non-ED peers, the data suggested that *CSEM* is an issue that disproportionately affects students who are economically disadvantaged compared to students who are not economically disadvantaged.

The results suggested that the tests in all states that reported data might be influenced by the out-of-school factors associated with being in the ED subgroup more than the in-school factors that influence achievement. The results suggested that inequity is being structured by faulty testing policy and score interpretation. Some students are being treated differently and

potentially not getting what they need as a result of proficiency miscategorization. The inequity is most severe in terms of who receives a high school diploma, who is allowed to take higher level courses, and who must be required to take low level basic skills instruction courses. Students in the ED subgroup are more likely to be miscategorized as *less than proficient* and more likely to experience negative consequences due to the miscategorization. Consequences can include lower lifetime income and shorter lifespan (Levin, 2009; Thomas, 2010).

CONCLUSIONS

Is *CSEM* a real concern for students? Yes, and even more so for students who are members of the ED group. According to the leadership of APA, AERA, NCME and individuals in the field of educational testing like Messick (1995, 1996) and Koretz (2008), the error inherent in the test results poses a negative construct validity issue because of the unintended consequences that it produces when SEA personnel do not report it and/or account for it through policy remedies. Construct validity issues are heightened when SEA personnel and others use the scores to make high-stakes decisions about students without considering error. Even a small amount of *CSEM* can have severe consequences for students when SEA personnel or school leaders simply require students to achieve a set cut-score to demonstrate proficiency (Koretz, 2008). The fact that one group whose membership includes some the nation's most fragile children is disproportionately affected negatively by policies that are known to lead to structured inequity is morally, ethically, and professionally troubling.

Because high school exams and *CSEM* are nationwide phenomena, we can be sure that hundreds of thousands of youth might be potentially affected negatively by inaction at the state and local levels to develop policy remedies and administrative practices aligned with standards and recommendations for appropriate testing practices. As stated in the *Standards*, "Measurement error reduces the usefulness of measures. … It reduces the confidence that can be placed in any single measurement" (p. 27).

POLICY RECOMMENDATIONS

One appropriate policy recommendation is for SEA personnel, policymakers, and school administrators to take a page out of the medical profession's handbook and to adhere to an education version of the Hippocratic oath, especially the oath of *do no harm*. Although education might have a similar oath, not all educators and education policy-makers seem to respect it. The time has come for school administrators to stop using the results from high stakes statewide high school exams to make high stakes decisions about children and to petition their state agencies to do the same. Until SEA personnel, policymakers, and all school administrators decide to protect children, put forth evidence based policies for appropriate practices, and *do no harm*, there should be a nationwide moratorium, through explicit policy language and law, on using such results for high-stakes decisions. At the very least, school leaders should adopt policies at the local level that limit the use of the test results for making high stakes decisions at the district level.

Because I see no signs of all school administrators or the policy-makers acting on my first set of recommendations, another approach to consider is to change the way *CSEM* is mitigated at the state level. One way is for states to keep their current number of testing opportunities but report all student scores with the *CSEM* band and award the highest score to the student (e.g., student's reported score plus the total amount of *CSEM* at the proficiency

cut-point). This increases the transparency of the process and helps overcome some score interpretation issues because the SEAs would recognize formally the *CSEM* on the individual score reports. This policy band-aid would help to ameliorate the potential negative social and educational consequences to students of not accounting for the *CSEM* when making decisions based on the scores. The score advantage should always go to the student in the high-stakes situation because of the inherent uncertainty and imprecision of the reported test results (APA, AERA, NCME, 1999). Including the *CSEM* in the student's score and awarding the score at the top end of the *CSEM*, along with multiple testing opportunities provides one procedural safeguard to lessen the unintended consequences due to *CSEM* precision issues recognizing, "Precision and consistency in measurement are always desirable. However, the need for precision increases as the consequences of decisions and interpretations grow in importance" (APA, AERA, NCME, 1999, p. 30).

CLOSING THOUGHTS

Children do not have a seat at the policy-making table. Policy is thrust upon them, and done to them, not with them. If those who make the policies and those who carry them out do not recognize or are unwilling to confront the potentially negative aspects of those policies and their actions, then children will be harmed, as they are every year. Perhaps, SEA personnel, policymakers, school leaders, and those who prepare them should be made to provide peer-reviewed, scientific evidence for their proposed policies and programs before those policies and programs are enacted. An agency like another Food and Drug Administration may be needed, but in this case an Education Protection Administration (EDPA), whose function is to review policies and programs through the lens of what's best for children and scientific evidence. The people who make and implement policies that mandate statewide testing and facilitate high-stakes decisions from the results need to consider inequities of our current system. The ends do not automatically justify the means. A "proficient" score, alone, does not have the empirical backing to support its reliability as the sole determiner of a student's academic performance.

REFERENCES

Achieve, Inc. (2008) *Closing the expectations gap*. Retrieved November 17, 2008 from http://www.achieve.org/files/50-state-2008-final02-25-08.pdf

Allensworth, E. M. (2005). Dropout rates after high-stakes testing in elementary school: A study of the contradictory effects of Chicago's efforts to end social promotion. *Educational Evaluation and Policy Analysis, 27*(4), 341–364.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

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

Bryk, A. S., & Hermanson, K. L. (1993). Educational indicator systems: Observations on their structure, interpretation, and use. *Review of Research in Education, 19*, 451–484.

Burch, P. (2005). The new educational privatization: Educational contracting and high-stakes accountability. *Teachers College Record* (online). Retrieved October 2, 2009, from http://www.tcrecord.org/content.asp?contentid=12259.

Cheeseman Day, J., & Newburger, E. C. (2002). The Big Payoff: Educational Attainment and Synthetic Estimates of Work-Life Earnings. *Current Populations Reports* P23–210. U. S. Census Bureau.

Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003) *State-Mandated Testing Programs on Teaching and Learning National Board on Testing and Public Policy*. Chestnut Hill, MH: Lynch School of Education, Boston College.

Education Commission of the States. (2008). *State notes: Exit exams*. Retrieved November 17, 2008 from http://mb2.ecs.org/reports/Report.aspx?id=1357

Fairtest.org. (2008). *Why graduation tests/exit exams fail to add value to high school diplomas*. Retrieved November 17, 2008 from http://www.fairtest.org/gradtestfactmay08.

Freedman, M. K. (2004). The fight for high standards. *Hoover Digest, 3*. Retrieved November 17, 2008 from http://www.hoover.org/publications/digest/3020841.html

Greene, J. P., & Winters, M. A. (2004). *Pushed out or pulled up: Exit exams and dropout rates in public high schools*. Manhattan Institute of Policy Research. Retrieved November 17, 2008 from http://www.manhattan-institute.org/html/ewp_05.htm

Haller, E. J., & Kleine, P. F. (2001). *Using educational research: A school administrator's guide*. New York: Longman.

Hanushek, E. A., & Welch, F. (2006). *Handbook of the economics of education*. St. Louis, MO: Elservier.

Harville, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practices, 10*(4) 181–189.

Harwell, M. R., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in educational research. *Educational Researcher, 39*, 120–131.

Heubert, J.P., & Hauser, R.M. (1999). High stakes testing for tracking and promotion. National Academy Press, Washington, D.C., p. 273-307.

Hursh, D. (2007). Policies assessing No Child Left Behind and the rise of neoliberal education. *American Educational Research Journal, 44*(3), 493–518.

Johnson, R. B. (2001). Toward a new classification of non-experimental quantitative research. *Educational Researcher, 30*(2), 3–13.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Laitsch, D. (2006). Assessment, high stakes, and alternative visions: Appropriate use of the right tools to leverage improvement. Retrieved on January 29, 2009 from http://epsl.asu.edu/epru/documents/EPSL-0611-222-EPRU.pdf

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Education Research Journal, 41*(4), 797–832.

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning*. New York: The Wallace Foundation.

Levin, H. (2009). The economic payoff to investing in educational justice. *Educational Researcher, 38*(1), 5–20.

Messick, S. (1995). Standards-based score interpretation: Establishing valid grounds for valid interpretations. *Proceedings on the joint conference of standard setting for large-scale assessments*, Sponsored by the National Assessment Governing Board and the National Center for Educational Statistics. Washington, DC: Government Printing Office.

Messick, S. (1996). *Technical issues in large-scale performance assessment*. Sponsored by the National Center for Educational Statistics. Washington, DC: Government Printing Office.

Neill, M. (1997). *Testing our children: A report card on state assessment systems*. FairTest.org. Retrieved on October 18, 2008 from http://www.fairtest.org/states/survey.htm or http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/0d/b3/a6.pdf.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Norris, S. P., Leighton, J. P., & Phillips, L. M. (2007). What's at stake in knowing the content and capabilities of children's minds. In R. Curren, (Ed.) *Philosophy of education: An anthology*. Hoboken, NJ: Wiley-Blackwell.

Ohanian, S. (2001). *One size fits few: The folly of educational standards*. Portsmouth, NH: Heinemann.

Penfield, R. D. (2010). Test-based grade retention: Does it stand up to professional standards for fair and appropriate test use? *Educational Researcher, 39*(2), 110–119.

Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low-achieving students to high-stakes testing. *Educational Evaluation and Policy Analysis, 23*(3), 197–227.

Stringfield, S. C., & Yakimowski-Srebnick, M. E. (2005). Promise, progress, problems, and paradoxes of three phases of schools accountability: A longitudinal case study of the Baltimore City Public Schools. *American Education Research Journal, 42*(1), 43–75.

Thomas, J. (2010, Jan. 4). Poverty, poor education shave years off life span. *Bloomberg Businessweek*. Retrieved from: http://www.businessweek.com/lifestyle/content/healthday/634465.html

Tienken, C. H. (2008). A descriptive study of the technical characteristics of the results of New Jersey's assessments of skills and knowledge in grades 3, 4, and GEPA. *New Jersey Journal of Supervision and Curriculum Development, 52,* 46–61.

Tienken, C. H. (2009). High School Exit Exams and Conditional Standard Error or Mismeasurement. *NCPEA Yearbook 2009*, p. 163–173.

United States Department of Education. (2010). *A Blueprint for Reform: The Elementary and Secondary Education Act* (ESEA) *Reauthorization*.

Vasquez Heilig, J., & Darling-Hammond, L. (2008). Accountability Texas-Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context. *Educational Evaluation and Policy Analysis, 30*(2), 75–110.

Weiss, C.H. (1998). *Evaluation* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Zwick, R. (2004). Rethinking the SAT: The future of standardizing testing in university admissions. New York: Rutledge.

Appendix A. Mean Scale Scores and Effect Size Differences on Statewide High School Exams in Mathematics (M) and Reading/Language Arts (LA) for Students Labeled Economically Disadvantaged and Not Disadvantaged

State	Proficiency Cut-Score LA/Math	LA/Math CSEM (SS)	ED. Scale Score, (SD), & N	Non-ED.Scale Score, (SD), & N	Effect Size Math/LA	Students Affected
AL	NA	-/-	No Data Reported	No Data Reported	NA	NA
AK (LA) (M)	NA	19 19	No Data Reported	No Data Reported	NA	412 937
AR (LA) (M)	NA	-/-	191 (21.7)12,793 203.9 (44.3) 16,669	206 (21.1) 16,870 231.2 (43.1) 16,781	0.69 0.62	NA NA
AZ (LA) (M)	NA	13 8	672 (93) 27,569 677 (91) 26,930	712 (70) 47,737 710 (67) 47,324	0.43 0.36	4906 1907
CA (LA) (M)	350 350	14 18	365.91 (34.8) 191,318 370.68 (36.7) 191,324	389.78 (36.1) 266,500 391.55 (38.3) 266,512	0.69 0.57	54,000 [3] 54,000
CO (LA) (M)	663 627	28 13	650.38 (62.1) 14,136 544.33 (71.6) 14,251	692.47 (56.9) 42,114 600.18 (69.2) 41,288	0.68 0.78	7594 4037
CT (LA) (M)	NA	-/-	211.40 (41.2), 10,349 218.20 (45.1) 10,320	254.50 (41.8) 31,432 264.30 (39.1) 31,374	1.05 1.02	NA NA
DE (LA) (M)	NA	10 10	501.82 (34.9) 2,418 518.66 (30.8) 2,595	525.17 (36.1) 5,975 542.52 (41.1) 6,138	0.67 0.77	773 977
FL (LA) (M)	300 300	19 8	282 (N/A) 69,044 313 (N/A) 68,748	321 (N/A) 114,368 336 (N/A) 114,003	NA NA	6,348 [3] 10,006
GA (LA) (M)	NA	9 5	No Data Reported	No Data Reported	NA	NA

HI (LA) (M)	NA	-/-	300.3 (40.5) 4,359 269 (41.2) 4,359	318 (38.5) 8,154 287 (41.4) 8,154	0.43 0.44	NA NA
IA (LA) (M)	NA	-/-	269.33 (41) 8,025 270.43 (37.7) 8,024	294 (39.6) 25,650 295 (36.5) 25,638	0.60 0.65	NA NA
ID (LA) (M)	NA	3.3 3.3	226 (N/A) 5,967 242 (N/A) 6,008	No Data Reported No Data Reported	NA NA	718 1226
IL (LA) (M)	155 156	4.03 6.75	148.39 (14) 35,361 147.62 (13.9) 35,392	159.69 (15.5) 95,385 160.60 (16) 95,413	0.81 0.94	NA NA
IN (LA) (M)	NA	-/-	549.5 (49.2) 27,700 577.9 (65) 27,700	583.3 (46.7) 51,046 621.3 (58.4) 51,046	0.69 0.67	NA NA
KS (LA) (M)	NA	-/-	70.68 (16.2) 9,033 48.53 (15.4) 9,069	80.3 (12.89) 22,597 55.40 (15.6) 22,503	0.59 0.45	NA NA
KY (LA) (M)	1040 1040	-/-	1039 (14.7) 21,775 1127 (19.2) 17,613	1048 (15.5) 27,414 1139 (21.6) 26,806	0.61 0.63	NA NA
LA (LA) (M)	299 305	15 15	292 (44) 21,503 309 (43) 21,497	314 (43) 20,091 337 (52) 20,093	0.50 0.65	NA NA
MA(LA) (M)	240 240	9 9	No Data Reported	No Data Reported	NA	7700 5600
MD LA) (M)	396 412	-/-	No Data Reported	No Data Reported	NA	NA
ME (LA) (M)	1142 1142	29 32	1134 (13.8) 3545 1136 (9.9) 3,695	1143 (14.4) 11,034 1142 (11.1) 11,175	0.65 0.61	7719 9360
MI (LA) (M)	1100 1100	89 88	1091 (32.7) 30,898 1078 (31.7) 30,694	1110 (30.5) 82,744 1098 (29.4) 82,540	0.58 0.63	19,371 25,794
MN(LA) (M)	1040 1140	14 12	1048.3 (13.7) 18,106 1129 (19) 15,605	1058.8 (12.4) 46,983 1144.6 (19.8) 46,832	0.77 0.82	5435 NA
MO(LA) (M)	NA	8 9	700.25 (35.9) 19,089 709.44 (48.7) 23,771	718 (34.8) 43,384 739 (46.7) 45,002	0.49 0.61	1,926 [3] 2,038
MS (LA) (M)	NA	-/-	646.44 (7.8) 13,936 649.44 (9.4) 14,527	652.9 (11.1) 14,118 654.9 (10.6) 14,683	0.83 0.58	NA NA
MT (LA) (M)	250 250	13 12	249.8 (35.1) 2,643 245.3 (27) 2,643	268.3 (31.4) 8,619 261.7 (27.4) 8,619	0.53 0.61	788 816
NC	NA	-/-	No Data Reported	No Data Reported	NA	NA
ND (LA) (M)	700 739	35 37	698.2 (29.8) 1,670 725.2 (41.4) 1,664	710 (27.3) 5,577 745.6 (36.5) 5,566	0.40 0.49	2500 1650
NE	NA	-/-	No Data Reported	No Data Reported	NA	NA
NH (LA) (M)	NA	-/-	1138 (N/A) 2,127 1129 (N/A) 2,106	1144 (N/A) 13,484 1134 (N/A) 13,440	NA NA	NA NA
NJ (LA) (M)	200 200	9 9	202.7 (N/A) 18,849 200.7 (N/A) 18,833	225.1 (N/A) 79,207 226.0 (N/A) 79,152	NA NA	9,500 [3] 9,500
NM*(LA9) (M)	NA	9 9	674.1 (33.7) 14,895 695.6 (34.1) 14,860	693.9 (35.2) 11,942 718.7 (40.2) 11,932	0.59 0.68	1664 1897

NY	NA	-/-	No Data Reported	No Data Reported	NA	NA
NV (LA) (M)	251 242	26 33	270 (60) 7,660 278 (57) 7,887	298 (57) 22,477 302 (56) 22,951	0.47 0.42	NA NA
OH (LA) (M)	400 400	8.59 10.02	No Data Reported	No Data Reported	NA	NA
OK[1] (LA) (M)	684 NA	-/-	No Data Reported	No Data Reported	NA	867 820
OR (LA) (M)	236 236	-/-	234.8 (8.7) 14,787 231.7 (9.3) 14,715	240.4 (8.9) 26,944 237.8 (10.2) 26,736	0.64 0.66	NA NA
PA (LA) (M)	1257 1304	54 49	1220 (252.6) 34,176 1210 (240.8) 34,231	1410 (266.5) 100,839 1390 (260.9) 100,906	0.75 0.74	12,223 9950
RI (LA) (M)	1140 1140	6 4	1138 (N/A) 3,367 1128 (N/A) 3,367	1145 (N/A) 8,294 1134 (N/A) 8,294	NA NA	1515 1982
SC (LA) (M)	200 200	5.6 5.5	218.26 (21) 20,319 215.19 (23) 20,682	234.76 (21) 26,495 232.27 (24) 26,727	0.79 0.74	995 [3] 1079
SD (LA) (M)	709 NA	-/-	720.35 (35.9) 1,771 715.35 (36.4) 1,778	738.51 (37.1) 6,875 734.47 (37.4) 6,890	0.51 0.53	NA NA
TN**(LA9) (M)	NA	-/-	515.5 (43.3) 23,605 526.4 (51.6) 12,871	543.7 (40) 40,489 554.2 (44) 21,119	0.65 0.54	NA NA
TX (LA) (M)	2100 2100	32 39	2217 (130.8) 130,407 2115 (172.6) 127,130	2296 (138.9) 167,764 2217 (197.1) 165,562	0.60 0.59	9,610 14,869
UT (LA) (M)	NA	-/-	161 (13.3) 9,383 157 (13.7) 8,062	168 (11.5) 27,283 164 (12.8) 32,588	0.53 0.51	NA NA
VA (LA) (M)	400 400	24 17	No Data Reported	No Data Reported	NA	1212 1445
VT (LA) (M)	NA	-/-	1139 (N/A) 1,489 1129 (N/A) 1,469	1145 (N/A) 5,751 1135 (N/A) 5,718	NA NA	NA NA
WA (LA) (M)	400 400	9 10	412.2 (31.3) 19,885 376.5 (38.2) 20,520	429.4 (30.5) 42,954 403.4 (39.7) 44,745	0.55 0.70	3623 5092
WI (LA) (M)	503 541	16 12	504.6 (61) 17,552 532.9 (48.4) 17,670	550.3 (55) 50,609 571.2 (44.7) 50,666	0.75 0.79	NA NA
WV	NA	-/-	No Data Reported	No Data Reported	NA	NA
WY[2](LA) (M)	159 148	5 5	151.5 (16) 826 143.1 (16.3) 894	157.7 (16.2) 2,827 149.5 (16.5) 3,167	0.39 0.40	NA NA

Note: All results come from the 2008 administration of each state's test in LA and M unless otherwise noted. All tests are either from Grade 10 or 11 unless otherwise noted.

[1]# of students affected estimated by using the number of students who were 2nd time test takers

[2]The mean scores for all learners are below the proficiency cut-points. This is due to a scoring error by Pearson that has not been corrected in the official score publications.

[3]# of students affected was estimated

*2007 data because 2008 were incomplete

**2006 data because 2007 and 2008 were incomplete

Citation: Tienken, C. H. (2011). *Structured inequity: The intersection of socio-economic status and the standard error of measurement of state mandated high school test results*. (p. 257-271). NCPEA Yearbook. Lancaster, PA: Proactive Publications.