High School Exit Exams Revisited: The Dangers of Relying on Null Hypothesis Statistical Testing when Considering Educational Policy

Author(s): ***Joe Dryden***

Affiliation: Texas Wesleyan University

2010

Abstract

The utilization of high school exit exams as a measure of accountability has produced considerable debate about their impact on completion rates. Every year, tens of thousands of students are denied a high school diploma despite completing all the requirements necessary for graduation except one; they cannot pass a state-mandated exit exam. Proponents of exit-level exams assert that these exams are necessary to ensure that graduates possess a minimum level of competency. Opponents assert that these exams have a negative impact on graduation rates especially among minority and lower income students. Research on this issue has produced mixed results when relying solely on traditional statistical measures. However, when the principles of effect sizes are examined it becomes clear that high school exit exams have a large, negative impact on graduation rates. Policy makers considering the use or implementation of high school exit exams should look beyond the dichotomous nature of traditional statistical measures to the more practical level of impact measured by effect sizes.

The Impact of High School Exit Exams and High Stakes Testing

Before we examine the limitations and dangers of relying on Null Hypothesis Statistical Testing when establishing educational policy, it is important that we understand the systemic impact of high stakes testing. What do we actually know about the impact of high stakes testing on children? First, data indicates that high stakes testing does produce improvements in student achievement in the areas tested. In 1994 for example, 75 percent of tenth grade students in Texas passed the Texas Assessment of Academic Skills (TAAS) reading test. In 2002, that number had risen to 94 percent. The results were even more impressive for math passage rates where 55 percent of tenth grade students passed the TAAS in 1994 compared to 92 percent in 2002 (Kosar, 2005). Standing alone, these data seem quite impressive, but at what cost, and how have these apparent improvements actually been achieved?

Educational success as measured by high stakes tests appear to be at the expense of other deserving areas of exploration and study. Corbett and Wilson (1991) found that high stakes testing narrowed the curriculum by placing more emphasis on basic skills while neglecting non-tested subjects. In many cases, the learning needs of students that have mastered the basic skills are neglected as teachers devote a substantial amount of time toward test preparation and the needs of struggling learners (Winebrenner & Espeland, 2001). A report from the Center for Educational Policy (2005) found that almost 50% of the districts surveyed reported that students are now spending more time focusing on reading and math. From one perspective, this should be celebrated as reading and math fluency are of critical importance; but these are not the only skills that schools should develop. By focusing a majority of our time and effort toward teaching reading and math, we neglect subjects such as, science, social studies, art, music, athletics, and fine arts. Torrance (1993) found that high-stakes tests resulted in teachers spending increased time with test preparation and test-taking strategies at the expense of other classroom duties and learning opportunities.

According to Abrams and Haney (2004), evaluating student performance and school performance on the basis of test-driven criteria has significant, negative, collateral impacts, including increases in dropout rates. In the 1980s in Texas, the graduation rates for low income students fell dramatically after the implementation of education reforms which instituted achievement testing as a method of accountability. Similar patterns have been observed in several other states such as New York and Florida (Losen, 2005). When faced with punitive measures for lack of educational improvement and a lack of adequate resources to serve students with learning difficulties, some school officials may resort to manipulative tactics to remove students with the greatest chance of failure.

Shepard (1991) identified a proximal link between the presence of high stakes testing and elevated retention rates. This phenomenon has been coined "academic red-shirting" and the results are disastrous. Alfie Kohn (2004) states:

> Some students are being forced to repeat a grade not because this is believed to be in their best interest, but because pressure for schools to show improved test results induces administrators to hold back potentially low scoring children the year before a key exam is administered. That way students in, say, tenth grade will be a year older with another year of test prep under their belts before they sit down and start bubbling ovals (pp. 94-95).

According to Walt Haney (2004), there were 13% more students in the ninth grade in 2000 than in the eighth grade in 1999. This practice of retaining students produces unintended, collateral damage especially when we take into consideration the statistics associated with grade retention on dropout rates. According to Jay Hubert (2003), grade retention is stronger predictor that a child will drop out of school than is socioeconomic status. Do we want to add this risk factor to the life experiences of thousands of children simply so they can pass one test?

Given the undeniable negative, collateral impact of high-stakes testing and high school exit exams, it seems logical that alternative routes to a diploma should be provided by any state that chooses to implement an exit-level testing requirement. Policy makers must realize that increasing standards produces both advantageous as well as disadvantageous consequences. Policy makers must realize that systems are interrelated and changes in one part of the system, can, and will produce an unintended ripple effect in other areas.

Implementing a policy, which requires students to pass an exit level exam will decrease completion rates unless the test is so easy that graduation rates are artificially held constant.

Another problem related to the utilization of high school exit exams lies in the fact that rigor of each exam varies considerably from state to state (Center for Educational Policy, 2008). This inequity means that a person living in one state may be denied a high school diploma while a student in a neighboring state, that either has no exit exam or one of low rigor, receives their high school diploma and all the privileges and opportunities that go with it. The National Center for Educational Statistics (2009) published a list of each state that requires an exit exam as a condition to receiving a high school diploma. This list includes a comparison of the subjects tested and the grade level at which each subject is tested. What is clear from the list is the tremendous variability in the rigor of each states exam not to mention that half of all states don't even require an exit exam.

Based upon cumulative exit-level data published on the website of the Texas Educational Agency, 24,733 students failed one or more sections of the exit-level Texas Assessment of Knowledge and Skills (TAKS) in 2005, 36,564 in 2006, 46,350 in 2007, 41,177 in 2008, and 36,568 in 2009 (Texas Education Agency, 2010). While each of these students can continue to take the exit-level test an unlimited number of times,

on average, only 35 to 40 percent those who retake the exit level exam pass. In addition, fewer and fewer students take the exit level with each proceeding year removed from high school. How can we allow these misguided policies, which are based upon contradictory data at best to continue to stand in the way of tens of thousands of children? We must consider alternate approaches toward determining who receives a high school diploma not just in Texas, but in any state considering the adoption of an exit-level exam.

Given the long list of collateral damage produced as a result of high stakes tests, any state considering their adoption and passage as a condition to receiving a high school diploma must recognize that higher standards will increase failure rates. Considering the impact on the life of a child that successfully navigates 13 years of school, yet cannot pass one portion of an exit-exam, it is imperative that look at multiple measures of statistical and practical significance when considering the use of exit level exams. Despite this assertion, strict reliance on traditional Null Hypothesis Statistical (NHST) still permeates the literature and dominates the decision making process. Policy makers should focus more on effect sizes and confidence intervals when considering research in the decision-making process.

The Use of Confidence Intervals and Effect Sizes

For the last several decades it has become clear that strict adherence to the principles of NHST has produced impediments to learning and unintended, negative consequences on a practical level (Anderson, Burnham, & Thompson, 2000; Daniel, 1998). Critics of NHST have even called for the American Psychological Association (APA) to consider banning NHST from all APA journals (Wilkinson & Task Force on Statistical Inference, 1999). Many researchers have advocated for a change away from NHST toward the use of effect sizes and confidence intervals (Byrd, 2007; Coe, 2002; Cummings and Finch, 2005; Nix & Barnette, 1998). In response to these concern, the American Psychological Association created a Task Force on Statistical Inference which recommended the use of confidence intervals in all research publications and described them as "the best reporting strategy" (APA, 2001, p.22). This language was incorporated in the APA Publication Manual fifth edition along with the suggestion that effect sizes be reported when available. The sixth edition of the APA Publication Manual, published in June of 2009, "stresses that NHST is but a starting point and that additional elements such as effect sizes, confidence intervals and extensive descriptions are needed to convey the most complete meaning of the results" (p. 33). The debate over the advantages and disadvantages of NHST has been going on for decades (Levin, 1998) and what has become clear is that change is a slow process What is also clear is that over reliance on NHST continues to impede the development of wise and sensible educational policy.

In the medical field, NHST was the standard for decades until critics pointed out that statisticians were making important health decisions, not medical professionals. During the 1970s and 1980s several prominent journals began to encourage the use of confidence interval (CI) and effect size (ES) reporting. In the field of ecology, where populations can in some cases be quite small, Type II errors, which are also referred to as false negatives, can have disastrous consequences (Fidler & Cumming 2007). Fidler and Cummings (2007) go on to state that "null hypothesis significance testing is widely misunderstood and misused and causes serious damage to research progress" (p. 441). Despite these legitimate concerns, and what Cohen (1994) called the inverse probability fallacy, where researchers are trying to disprove a null hypothesis, NHST remains the primary way researchers draw conclusions about educational practices as opposed to the use of CIs and ESs (Fidler & Cumming 2007).

Coe (2002) argues that not only are effect sizes easy to calculate, but they clearly show the magnitude of the difference between two or more groups. ESs allow researchers to quantify the magnitude of the effectiveness of intervention strategies or policy decisions. In addition, the use of ESs allow researchers to

go beyond answering whether something works to a more mature and practically significant understanding of how well an intervention works and in which context. Fidler and Cumming (2007) agree by describing NHST as dichotomous; either an intervention works or it does not when most researchers recognize that the answer is not always so clear. Coe (2002) goes on to say that the use of effect sizes are significantly underutilized in quantitative research.

The explanation for the limited use of effect sizes when reporting quantitative research involves several factors. First, ESs are scarcely mentioned in most statistical textbooks, rarely taught in research courses and often misunderstood by practicing researchers. Further compounding the problem is the observation that even when CIs and ESs are reported, they are rarely discussed or interpreted correctly or relied upon for the purpose of drawing conclusion, making recommendations or developing policy (Cummings, Fidler, Leonard, Kalinowski, Christiansen, Kleinig, Lo, McMenamin, & Wilson, 2007). Cummings et al. (2007) go on to assert that journal editors and institutional programs must take an active role by supporting the change from the traditional NHST approach to the use of CIs and ESs in all research applications. Chudowsky and Gayler (2003) assert that research on the impact of high school exit exams is limited and as a result policy decisions are being made without sufficient or in many cases misleading data.

Contradictory and Misleading Statistical Data

Greene and Winters (2004) assert that based solely on intuition, it makes sense that increased standards would increase dropout rates; however, they go on to assert that the evidence is not conclusive. Greene and Winters (2004) analyzed the work of several other researchers looking at the same question and noted that the analysis of most other researchers relied solely on "the dichotomous measure of whether states made gains or losses relative to the national average" (p. 2). Greene and Winters (2004) described additional studies that produced conflicting findings possibly as the result of the data used to measure graduation rates. Amrein and Berliner (2002) studied whether states that implemented exit exams observed increased dropout rates, or increased numbers of students pursuing a GED instead of a high school diploma. Their findings indicated that more than half of states that adopted high school exit exams observed increased dropout rates higher than the national average. Contradictory findings were reported by Carnoy and Loeb (2003) who also studied the impact of high stakes testing on graduation rates. Carnoy and Loeb (2003) found no relationship between the rigor of a state's exit-level exam and its high school graduation rates. Greene and Winters (2004) stated that the measures used by previous researchers failed to examine the magnitude of the effect, a problem that they themselves appear to repeat. With these contradictory findings how can policy makers make informed decisions? The answer lies in the reporting of more than just statistical significance manufactured through NHST. Policy makers must also utilize effect sizes and confidence intervals as evidence of the practical impact of policy decisions.

Greene and Winters (2004) used two different "highly respected graduation rate calculations" (p. 4) to measure the impact of high school exit exams. The first method developed by Greene "divides the number of diplomas awarded by a state in a given year by the estimated number of students who entered ninth grade four years earlier" (p. 4). The second method takes "the number of diplomas awarded by public schools in a given state by the number of 17-year-olds in a state's population during that year according to the U.S Census" (p. 4). After controlling for school funding and student teacher ratios, Greene and Winters (2004) utilized a fixed effects regression model to analyze the data.

Under the Greene method for calculating graduation rates, the p value was .423 and under the Census method, the p value was .143; neither of which are statistically significant. Consequently, Greene and Winters (2004) reported that the implementation of high school exit exams have no statistically significant effect on high school graduation rates. Given the reputation of Greene and Winters (2004) and that of the Manhattan Institute for Policy Research, several states relied on these findings as evidence that

the implementation of high school exit exams would not have an effect on graduation rates. Fortunately, or unfortunately depending on whether you are a proponent or opponent of high school exit exams, Greene and Winters (2004) also calculated and reported effect sizes for each method. Under the Greene method, the implementation of high school exit exams had an effect size of -.76437. Under the Census method, the implementation of high school exit exams had an effect size of -1.11624. What do these effect sizes mean in terms of practical significance?

As a rule of thumb, an effect size of 0.2 is considered to be small, 0.5 is considered to be a medium effect size and 0.8 is considered to be a large effect size (Cohen, 1969). According to Coe (2002), effect sizes are equivalent to a Z score in a Normal distribution curve. If the measured effect size for an intervention is 1.0 for the experimental group, the average participant from that group would score higher than 84% of the members from the control group. If an intervention produced an effect size of -1.0, the average participant would drop one standard deviation below the mean to the 16th percentile. Applying the effect sizes reported by Greene and Winters (2004), (-.76437 and -1.11624) means that a state which was at the 50th percentile in terms of graduation rates prior to the implementation of a high school exit exam would fall to somewhere between the 14th and 22nd percentile. While these results may not be considered significant under the traditional NHST approach, it is clear that the implementation of a high school exit exam has a large negative effect on graduation rates.

Alternatives to the Status Quo

Every year thousands of students are denied a high school diploma despite completing all the requirements for graduation except one; passing a state mandated exit exam. This being the case, it is imperative that those entrusted with the authority to make educational policy look beyond the dichotomous, either or approach of NHST and consider the use of effect sizes and confidence intervals in the decision making process. The yes or no world of Null Hypothesis Statistical Testing (NHST) is simply not appropriate as the only measure of impact. Basing educational policy on all measures of statistical significance including ESs and CIs as opposed to just one dichotomous measure should allow policy makers to produce policies that are fair, just and sensible.

In the alternative, if a state decides to implement an exit-level testing requirement they should seriously consider providing an alternative path toward graduation. In 2008, 24 states required high school exit exams. Of these 24 states, 21 provided an alternative pathway, which allowed students to receive a diploma despite not passing every section of their states exit exam (Center for Educational Policy, 2008). Texas is one of the few states that does not offer an alternative pathway to graduation. In response to this near-sighted policy omission, there is a movement growing across the state led by former Commissioner Mike Moses to offer multiple pathways to graduation (Raise Your Hand Texas, 2010).

Texas should follow the example set by states such as North Carolina, Maryland, Indiana and Georgia which offer evidenced-based waivers or an alternative documentation processes where high school diplomas can be awarded based upon the good faith efforts of students, documented proficiency, and the recommendations of school officials. Even if a student cannot pass the exit-level exam, he or she could still have a chance at obtaining a minimum-level high school diploma under the following conditions: (a) the student has an attendance rate of 95% during their senior year, (b) the students maintains a C average or better in all classes during their senior year, (c) The student attends 90% of all exit-level exam tutorial sessions offered by the school, (d) the student takes advantage of every retake opportunity and (e) the student obtains a letter of recommendation for his or her principal based upon documented proficiencies.

Can we find a balance to maintain the commitment to promote and demand high levels of performance for high school graduates while maintaining the dignity of all children? Will we recognize that higher

standards come with a price; lower high school completion rates? Can we continue to carry the mantra of No Child Left Behind then leave thousands behind within sight of the finish line? The education of children is perhaps the most important function of any government and as such all educational policy should be based upon more than simply the dichotomous nature of NHST. We must consider the practical significance as measured by effect sizes and confidence intervals.

References

Abrams, L., & Haney, W. (2004). Accountability and the grade 9 to 10 transition: The impact on attrition and retention rates. In G. Orfield (Ed.), *Dropouts in America: Confronting the graduation rate crisis* (pp. 181-206). Cambridge, MA: Harvard Education Publishing Group.

American Psychological Association. (2001). Publication Manual of the American Psychological Association (5th ed.). Washington, DC.

American Psychological Association. (2010). Publication Manual of the American Psychological Association (6th ed.). Washington, DC.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*, 18. Retrieved from http://epaa.asu.edu/ojs/article/viewFile/297/423

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence and an alternative. *Journal of Wildlife Management*, *64*, 912-923.

Byrd, J. (2007) A call for statistical reform in EAQ. *Educational Administration Quarterly*, *43*(3), 381-391.

Carnoy, M., & Loeb, S. (2003). Does external accountability affect student outcomes? A cross-state analysis. *Education Evaluation and Policy Analysis*, *24*, 4, 305-331.

Center for Educational Policy. (2005, March). From the capital to the classroom: Year 3 of the no child left behind act. Retrieved October 20, 2006, from http://www.cepdc.org/pubs/nclby3/press/cep-nclby3_21mar2005.pdf.

Center for Educational Policy, (2008). *State high school exit exams: A move toward end of course exams*. Retrieved from http://www.cep-dc.org

Chudowsky, N., & Gayler, K. (2003). Effects of high school exit exams on drop out rates: Summary of a panel discussion. Center on Educational Policy. Retrieved from http://www.manhattan-institute.org/pdf/ewp_05.pdf

Coe, R. (2002, September). *It's the effect size stupid: What is effect size and why is it important?* Paper presented at the British Educational Research Association conference. University of Exeter, England. Retrieved from http://www.leeds.ac.uk/educol/documnets/00002182.htm.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press.

Corbett, H. D., & Wilson, B. (1991). *Testing, reform, and rebellion*. Norwood, NJ: Ablex.

Cummings, G., Fidler, F., Leonard, M., Kalinowski, p., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology: is anything changing? *Psychological Science*, *18(*3), 230-232.

Cummings, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170-180.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for editorial policies of educational journals. *Research in the Schools*, *5*(2), 23-32.

Fidler, F., & Cumming, G. (2007). Lessons learned from statistical testing in other disciplines. *Psychology in the Schools*, *44*(5), 441-449.

Greene, J. P., & Winters, M. A. (2004). Pushed out or pulled up? Exit exams and dropout rates in public high schools. Retrieved from http://www.manhattaninstitute.org/pdf/ewp_05.pdf

Haney, W. (2004). *The education pipeline in the United States*, 1970-2000. Boston, MA: National Board on Educational Testing and Policy.

Hubert, J. P. (2003). First do no harm. *Educational Leadership, 60,* 26-30.

Johnson, T., Boyden, J. E., & Pittz, W. J. (2001). Racial profiling and punishment in U.S. public schools: How zero tolerance and high stakes testing subvert academic excellence and racial equity. Retrieved from http://www.manhattan-institute.org/pdf/ewp_05.pdf

Kohn, A. (2004). NCLB and the effort to privatize public education. In Meier D. &Wood, G. (Eds.) *Many children left behind: How the no child left behind act is damaging our children and our schools* (pp. 79-97) Boston MA: Beacon Press.

Kosar, K. R. (2005). *Failing grades: The federal politics of educational standards*. Boulder, CO: Lynne Rienner Publishing.

Levin, J. R. (1998). What if there were no more bickering about statistical significance test? *Research in the Schools*, *5*(2) 43-53.

Losen, D. J. (2005). Graduation rate accountability under the no child left behind act. In Sunderman, G., Kim, J., & Orfield G. (Eds.), *NCLB meets school realities*. (pp. 105-120). Thousand Oaks, CA: Corwin Press.

National Center for Educational Statistics (2009). State high school exit exams, by exam characteristics and state: 2008-09. Retrieved from http://nces.ed.gov/programs/statereform/tab5_5.asp

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, *5*(2), 3-14.

Raise Your Hand Texas, (2010). Multiple pathways to high school graduation. Retrieved from http://www.raiseyourhandtexas.org/site/PageServer?pagename=what_legagenda

Shepard, L. A. (1991) Will national tests improve student learning? *Phi Delta Kappan, 73 (3)*, 232-38.

Texas Education Agency, (2010). Academic excellence indicator system. Retrieved from http://ritter.tea.state.tx.us/perfreport/aeis/index.html

Torrance, H. (1993). Combining measurement-driven instruction with authentic assessment: Some initial observations of national assessment in England and Whales. *Educational Evaluation and Policy Analysis*, 15(1), 81-90.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Winebrenner, S., & Espeland, P. (2001). *Teaching Gifted Kids in the Regular Classroom: Strategies and Techniques Every Teacher Can Use to Meet the Academic Needs of the Gifted and Talented*. Minneapolis, MN: Free Spirit Publishing.