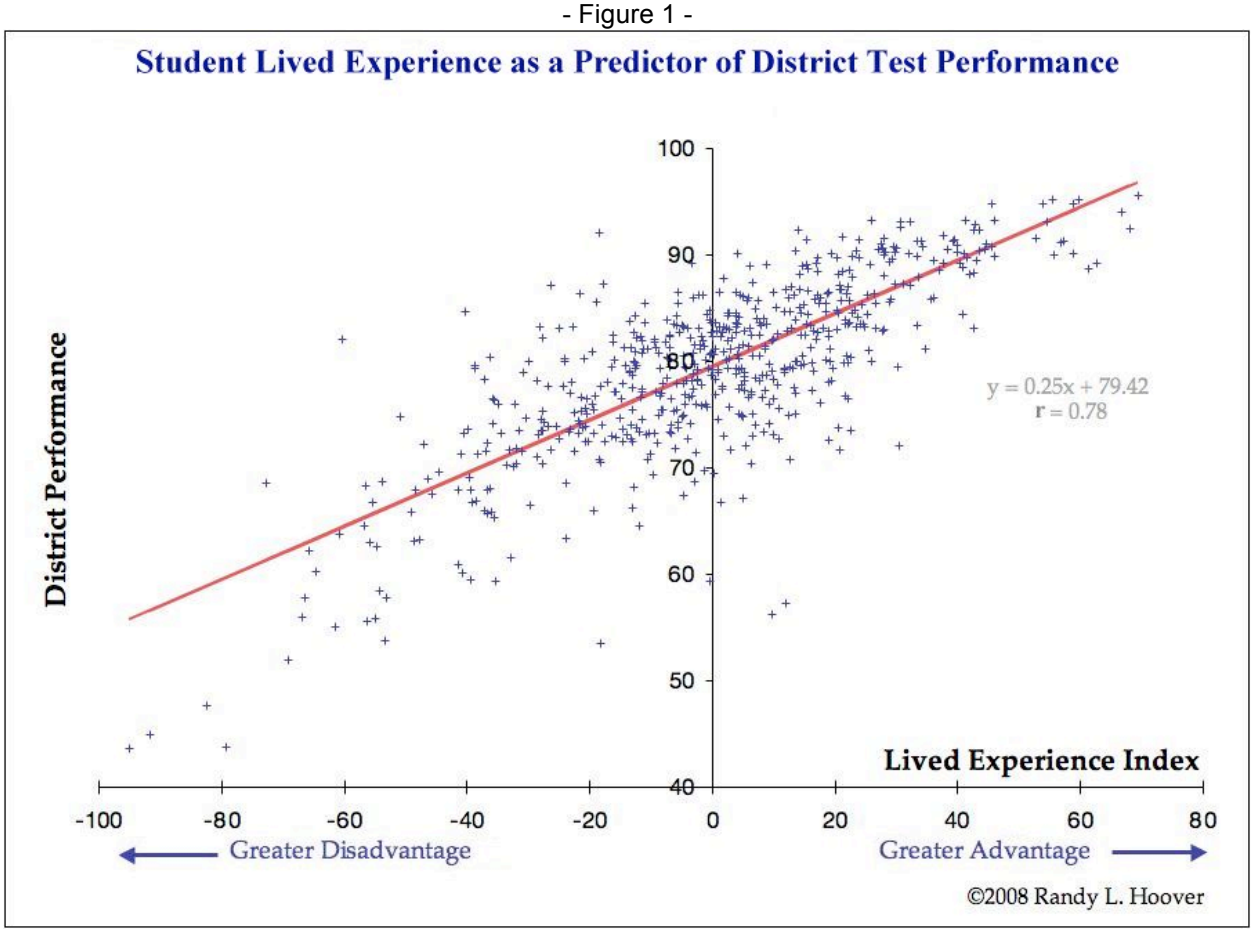
Some Research Insights about the Validity of Standardized Tests in Ohio[1]

Randy L. Hoover, PhD (2014)

Mathematical analysis of standardized tests and the associated metrics used to produce claims and conclusions about student achievement and educator performance are complex yet accurate in their determinations. The most vital and immediate determination to be made about any standardized test is its *validity*. *Test Validity refers to the degree to which a particular test accurately measures that which it claims to measure*. Test validity is only meaningful in terms of how the test is used and what it is used for. In other words, test validity is a formal examination to determine the degree to which a test is appropriate and accurate in serving what it is used for. Formal examination reveals that the assumption of Ohio's tests being valid is false. *The Ohio achievement tests fail to meet the accepted standards for test validity*[1] [2].

- Figure 1 -



The essence of the matter of the validity failure of Ohio's tests is shown in *Figure 1*. The graph shows the extremely high and statistically significant correlation between test performance and the nature of the life experience of those taking the tests. Indeed, with a correlation where r = 0.78 out of the highest possible correlation where r = 1.00, the tests are shown to be extremely sensitive in measuring life experience rather than academic achievement. *These results clearly and emphatically validate what every experienced educator knows intuitively—the conditions in which a child grows up (forces and factors such as family, neighborhood, income, nutrition, health, and psychological stability among others) are overwhelmingly the primary determinant variables of academic achievement.*

Hart and Risley (1995) give us tremendous insight into why the living conditions of the child factor so greatly into standardized test performance in an excerpt of their book[3] published by The American

[1] For much more detailed explanation and research on Ohio achievement test validity, hit the Research button at the top or go directly to http://teacher-advocate.com/research .

Federation of Teachers. Their research focuses on how differently children experience language in the home setting and on how differently they experience emotional support and criticism across socio-economic levels. Because the standardized tests are so language bound, Hart and Risley present sound evidence as to why children of different socio-economic groups exhibit such differential test performance.

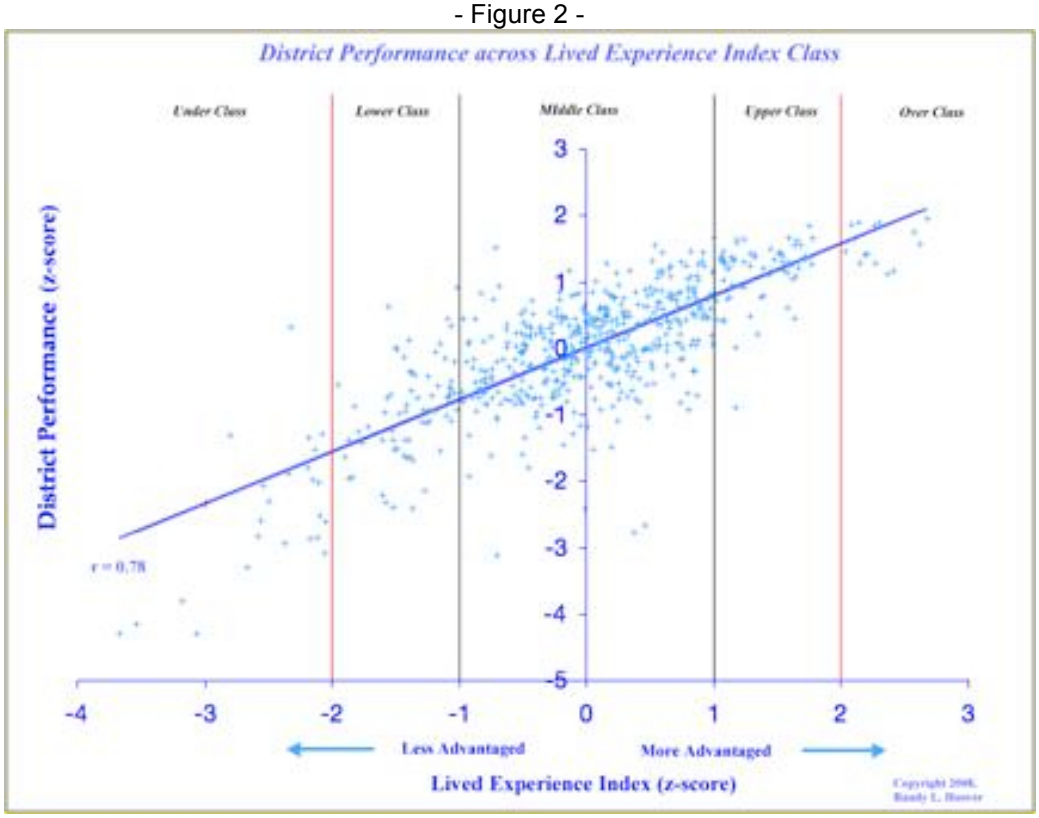
Figure 1 is derived from my second study of test performance in Ohio (Hoover, 2008). This particular study examined 609 of the 611 Ohio school districts on all sections of the 2007 third-grade, fourth-grade, fifth-grade, sixth-grade, seventh-grade, and eighth-grade Ohio Achievement Tests, and the Ohio Graduation Test (*Table 1*). Therefore, the research analysis used 23 sets of test data for each of the 609

- TABLE 1 -
2007 Grade-Level and Subject-Area Test Data Sources

Grade Level	Reading	Mathematics	Writing	Social Studies	Science
3rd Grade	X	X			
4th Grade	X	X	X		
5th Grade	X	X		X	X
6th Grade	X	X			
7th Grade	X	X	X		
8th Grade	X	X		X	X
OGT	X	X	X	X	X

school districts--a total of more than 14,000 data cells representing Ohio school district performance. The lived experience indicator was derived from the percent of students enrolled in the federal lunch program, average family income, and number of single-parent families of each district. I use this as the example simply because it is more recent than the first study done in 1999. The 1999 study of 593 districts actually showed a slightly higher correlation (r = 0.80). The critical point upon which Ohio's validity failure turns comes from the powerful association of out-of-school lived experience and test performance in terms of economic class as is seen in *Figure 2*.

- Figure 2 -



The scatter plot alone clearly shows the test performance as a function of the socio-economic conditions lived by the students. It also graphically shows the bias of the tests by *group membership*, a term used in the *Standards for Educational and Psychological Testing*[4] in reference to typical external factors such as race, class, gender, ethnicity, disability, or lifestyle. More specifically, the scatter plot and its associated

mathematical analysis illustrate that there is a very strong correlation of test performance to the combined effects the Lived Experience Index of median income, single parent family, and eligibility for federal lunch subsidy. The analysis provides strong evidence that group membership effects such as this should act as a significant caveat about claims of test validity when the evidence clearly shows a strong relationship of test scores to categories of external variables.

Establishing test validity requires the application of rigorous psychometric procedures to produce public evidence that 1) *The test questions cover a fair and representative sample of the academic standards it is intended to test* (internal validity), and 2) *The test content and language does not unfairly disadvantage any students because of particular group membership* (Standards for Educational and Psychological Testing, 1999; American Institutes for Research, 2008[5]) *related to ethnicity, race, class, or gender among others*. The first item above intends to assure that what is being tested accurately represents the breadth and depth of what the test takers have allegedly had the opportunity to learn based upon Ohio's established academic standards. The second item is intended to assure that no test bias exists due to student lived experience in terms of accepted demographic realities of race, class, gender, or ethnicity.

Ohio does nothing about item two above—the state does not address bias as related to test validity. Ohio claims that its "Fairness & Sensitivity Committee (FSC) is charged with ensuring that the test content does not unfairly disadvantage students because of group membership" (American Institutes for Research, 2008, p. 7). However, the FSC does no statistical analysis for external validity whatsoever. It merely discusses fairness in terms of item clarity and in terms of the distribution of minority names and references in test items. For Ohio or AIR, Ohio's technical test analysis vendor, to claim this committee addresses validity in terms of significant bias across race, class, gender, ethnicity, or disability is a complete misrepresentation of the committee calculated to hide the reality that the tests are demonstrably not valid psychometrically. In 2000, State Superintendent Susan Zelman attempted to dismiss my research using the argument that the Fairness & Sensitivity Committee guaranteed that bias across group membership could not happen. The 2000 study data and analysis proved her incorrect.

There is also a very frustrating irony in what the achievement gaps reveal that is directly connected to my research findings on Ohio test (in)validity that is never given as an interpretation by the reformists. Both research studies (Hoover, 2000;2008) clearly confirm that the achievement tests measure the lived experience of the student, not the academic achievement resulting from school and teacher effects. Given these findings, we would predict performance gaps across groups with significantly different cultural, home, and neighborhood environments. Indeed, that is precisely what the achievement gaps reveal as they provide additional evidence for the Ohio validity findings of the two studies. Yet in no reformist discussions that I am aware of is the nature of the achievement tests suggested as the cause of the much publicized achievement gaps.

Summary

Formal analysis of validity mathematically links a specific test to its specific use to determine the degree of integrity the test results have for use as accurate representations of individual test performance. *Therefore, any high-stakes reward or punishment based upon Ohio's tests is indefensible scientifically and outrageous morally.* Denial of graduation based upon an OGT standardized test score, for example, is arbitrary and wrong from any reasoned perspective. There is also a very strong argument that it is a violation of the 14th amendment right to due process.

Further, this factual reality of Ohio's test validity failure subsequently and directly destroys the integrity of any and all its uses as data inputs for additional analysis such as Ohio's building or school ratings and value-added performance claims. The inescapable conclusion is, therefore, that Ohio's teacher and administrator evaluation system is in no way, manner, shape or form indicative of professional effectiveness. We need to understand that any educator evaluation element based upon the test scores is a sham beyond any doubt. It is pseudo accountability in extreme.

At the risk of being repetitious, but with the hope of greater understanding of the big picture view, I offer a paragraph from a chapter I recently wrote for the second edition (in press, 2014) of *The phenomenon of Obama and the agenda for education: Can hope audaciously trump neoliberalism?*[6]

> Test validity is the baseline psychometric requirement for any statistical procedure or metric that is based upon test scores. If the test data analyzed by the metrics are not demonstrably valid, nothing that flows from the system of metrics can be considered worthy of belief—credible. No matter how robust the system of metrics, the lack of test validity renders the outputs of the system worthless for making any defensible inferences about performance or for making any rational judgments for evidence-based recommendations for reform. The old adage of "garbage in, garbage out" plays out fully and completely in the end results. Such it is with. . . school reform metrics. However, if the machine is deliberately designed to advance a hidden agenda, then obfuscation of the accepted standards and principles of tests and measurements serves to create machine outputs that function as *false proxies* for school reform to be fed to the public in order to gain popular support… (Hoover, 2014; Godin, 2012).

---

[1] Hoover, R. (2000). *Forces and factors affecting Ohio proficiency test performance: A study of 593 Ohio school districts*. Retrieved from http://teacher-advocate/research

[2] Hoover, R. (2008). *A Re-examination of forces and factors affecting Ohio school district OAT and OGT performance.* Retrieved from http://teacher-advocate/research

[3] Hart, B., & Risley, T. (1995). The early catastrophe: The 30 million word gap by age 3. Excerpted from *Meaningful differences in the everyday experiences of young American children*. Brooks Publishing. Retrieved from http://www.aft.org/newspubs/periodicals/ae/spring2003/hart.cfm

[4] Standards for educational and psychological testing.

[5] American Institutes for Research. (2008). *Validity evidence based on internal structure: Examination of the factor structure of the Ohio achievement tests*. Report TR 2008-02, to the Ohio Department of Education. Retrieved from http://education.ohio.gov/getattachment/Topics/Testing/Ohio-Graduation-Test-OGT/TR-2008-2-OGT-Validity-Study-2007.pdf.aspx

[6] Hoover, R. (2014). The neoliberal metrics of the false proxy and pseudoaccountability. In Carr, P. & Porfilio, B. (Eds.), *The phenomenon of Obama and the agenda for education: Can hope audaciously trump neoliberalism?* (2nd Edition) Charlotte, NC: Information Age Publishing.

Godin, S. (2012). Seth's Blog http://sethgodin.typepad.com/seths_blog/2012/11/avoiding-the-false-proxy-trap.html