GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE

DOCUMENT REVIEWED:	"Portfolios: A Backward Step in School Accountability"
AUTHOR:	Robert Holland
PUBLISHER/THINK TANK:	Lexington Institute
DOCUMENT RELEASE DATE:	September 2007
REVIEW DATE:	September 19, 2007
REVIEWER:	William J. Mathis
E-MAIL ADDRESS:	WMathis@sover.net
PHONE NUMBER:	(802) 247-5757

Summary of Review

This self-described "research study" and following press release are intended to influence the debate over the direction of the reauthorization of NCLB, offering a defense of the current test-based accountability system against the inclusion of "multiple measures." The report presents a review of the research on portfolios in large-scale school accountability systems, concludes that portfolio assessment is severely flawed, and then characterizes portfolios as a proxy for all non-test-based measures of student performance. The report has several glaring weaknesses, however. The literature review cherry-picks two studies, both conducted 13 years ago and, on the basis of those studies, concludes that portfolios are not reliable and are too expensive for large scale accountability systems. Yet other large scale studies of portfolios – some of which are discussed in one of the two studies that the report itself relies on – come to different conclusions but are not examined or even mentioned. An even bigger problem with this new report (which is repeated in the press release), however, is the author's decision to present portfolios as somehow representative of all non-test-based measures of student performance – which they clearly are not. This results in a document that is of little value for research or policy development.

Review

I. INTRODUCTION

The definition of school accountability in the reauthorization of the Elementary and Secondary Education Act (also known as No Child Left Behind [NCLB]) will have enormous impact on the direction of public education in the United States. This federal law establishes a forceful evaluation system that encompasses the vast majority of public schools in the nation. The NCLB law provides less than five percent of total education spending, but it exerts a disproportionate influence on policy due in large part to a provision requiring states to have uniform within-state accountability systems.

In the study under review – "Portfolios: A Backward Step in School Accountability" – and in a following press release on the study[1], the Lexington Institute argues that the federal government must retain a standardized-test based system and not use "multiple measures" of school performance, particularly portfolios.

Just days before the 2007 Labor Day recess, U. S. House of Representatives Education Chairman George Miller released his summary of proposed changes for the law's reauthorization. This draft would allow a state to use "multiple indicators" (the equivalent of Lexington's "multiple measures") of school achievement beyond statewide reading, mathematics, and science examinations. These measures could count for as much as 15% of an elementary school's annual performance target and up to 25% of a high school's score. Significant for this analysis, Representative Miller's summary does not use the term "portfolio" or even use the term as an example of multiple methods.[2]

The accountability system that NLCB originally put in place is almost exclusively based on standardized test results in reading and mathematics, although science examinations in two grades are phasing in. A single outside "academic indicator" is already a part of the equation. In most cases, this academic indicator is attendance rate or graduation rate.

Not wishing to be labeled in the media as a failing school or to be subjected to ever-increasing sanctions, schools have responded to the federal accountability system with various approaches designed to improve their NCLB results. The Center on Education Policy has found, for example, that the current emphasis on tests has led to the narrowing of the curriculum at the expense of other subject matters such as the arts, sciences and social studies.[3] In addition, Nichols and Berliner have documented many instances whereby NLCB has brought on the "corruption" of test scores' validity and the distortion of teaching and learning.[4]

Contending that schools have broader purposes than measured on multiple choice tests, an extensive coalition of education and non-educational groups say that multiple indicators are required to validly measure the outcomes of education in light of its diverse purposes in a democratic society.[5] The movement to include multiple indicators is also a response to concerns about the unintended negative consequences of the current law's overwhelming reliance on test scores.

This perspective has been countered by those arguing that a system based overwhelmingly on standardized tests provides essential reform pressure on schools. Multiple indicators and the use of assessment

methods such as portfolios would dilute the focus and relieve the pressure needed to force school reform, they contend.[6] The new report from the Lexington Institute is part of that pushback.

II. FINDINGS AND CONCLUSIONS OF THE REPORT

Listed as a "research study" on the institute's web site, this ten-page report's executive summary begins with a statement about the National Education Association's lobbyists, decries portfolio assessment as the most notable "multiple measure," declares it a "huge flop," and concludes with, "The question is why anyone sincerely interested in holding schools accountable for results would want to revive such a failed method of assessment."[7]

The report's main body begins with a transcript of an exchange between Chairman Miller and a reporter at the National Press Club. The Lexington report's author, Robert Holland, characterizes Miller as "backpedaling" on a question about using portfolios for non-English speaking students (p. 2).

This segues into the author's elucidation of the "NEA's Push for Portfolios" (p. 3). As evidence for this push, the author quotes a third level sub-point in the NEA's legislative priorities document, which mentions "performance or portfolio assessments" among a list of potential multiple indicators that go beyond standardized testing.[8] From this listing, the author concludes:

> . . . it is reasonable to suspect that the 3.2 million-member union wants more subjective forms of testing in order to conceal the reality that schools are failing to teach children . . . (p. 3)

Under the subsection title, "Concessions to the NEA?" Holland presents as evidence House Education Chairman Miller's use of the term "multiple measures," along with his statement that the nation needs higher-level skills and problem-solving skills.

A short discussion is then provided which takes to task as outcomes-based education Marc Tucker's "New Commission on the Skills of the American Workforce" report (*Tough Choices or Tough Times*) (p. 4). This is largely a diversion, since nothing in this section addresses either portfolios or multiple indicators.

In the second third of the Lexington report, the validity and practicality of portfolio assessment is addressed. The argument is grounded in two studies: one from Vermont and one from Kentucky. These are cited as evidence that portfolios are too unreliable to be used for school accountability. Both studies were conducted in 1994. The Vermont study was peer reviewed, but the citation provided in the report is only to the article's abstract. According to the Lexington report, the full article says there were large variations in scoring between teachers and that the costs were high.[9] The lengthy and technical Kentucky research report (which was not peer reviewed) is summarized in the Lexington report through a block quote from a secondary analysis conducted by the Pacific Research Institute (PRI), which, like the Lexington Institute, is a free-market think tank. (PRI is based in San Francisco.)

The last third of the paper revolves around another lengthy block quote, this time from the Kentucky report itself, intended to illustrate the subjective nature of portfolio assessment. An email from a retired Louisville professor is then provided, asserting that portfolios are popular with "radical constructivists" (p. 7). Then, a short section on

the excessive expense of scoring portfolios is presented (p. 8).

Echoing the executive summary, Holland's conclusion opines, "It is difficult to comprehend why any consideration is being given to reviving portfolio assessment as a way to gauge the effectiveness of No Child Left Behind …." (p. 9). Among the listed shortcomings of portfolio assessment are unreliability, differences in implementation between sites, differences in the difficulty of student assignments, and costs.

III. RATIONALES SUPPORTING THE FINDINGS AND CONCLUSIONS OF THE REPORT

The Lexington "research study" departs from traditional rationales and protocols.

The explicit conclusion is that standardized tests are "the best value in terms of reliability, accuracy, ability to generalize the results, ease of scoring and costs" (p.9) For this reason, the National Education Association's endorsement (along with that of other groups) of "multiple measures" in the NCLB accountability system should be rejected. To support this conclusion, portfolios are used as a proxy for all other possible indicators. The virtues and vices of portfolio assessments are implicitly extended to all non-test based measures. A justification is not offered for using portfolios as a stand-in for all multiple indicators.

The broader conclusion – that multiple indicators are untrustworthy, and NCLB should remain as a test-based accountability system – is supported only by the two referenced, 13-year-old studies.

Uses of rhetorical questions ("... Why anyone interested in . . . would want to revive such a failed method. . ."), loaded language ("a huge flop"), debatable interpretations of political statements (by Chairman Miller and by NEA President Reg Weaver), and departures from the topic (a criticism of outcomes-based workforce development proposals), all cloud the clarity, rationale and coherence of the report.

IV. THE REPORT'S USE OF RESEARCH LITERATURE

As noted, the report uses only two dated primary research sources. The Hambleton *et al* study[10] is a massive but not peer-reviewed study done for the Kentucky legislature. The Koretz *et al* citation is to an 11-page, peer-reviewed article about Vermont. As noted earlier, the inconsistency of writing prompts, scoring procedures and costs were the primary findings of both of these studies.[11]

The Lexington report provides its readers with only limited information about these studies. In fact, the large Kentucky study is mainly presented through a secondary source – an analysis written by Lance Izumi of the free-market Pacific Research Institute.[12] The report reproduces a passage from Izumi's analysis that laments the difficulty of equating portfolio examinations from year to year, the lack of implementation controls, and the failure of the Kentucky program to show comparable test score gains on the National Assessment of Education Progress.

Although an abundant literature exists in portfolio assessment (particularly for improvement of instruction), none was cited. This is particularly troubling in that the Koretz *et al* paper describes two other large-scale, major studies, both of which Koretz *et al* found to produce higher reliability from one rater to another, even though one of the studies used inferential terms (e.g. – "growth.").[13] Likewise, although educational

assessment and educational accountability are the center of interest for a large number of periodicals and professional groups, none of these resources were cited other than the Vermont and Kentucky studies. Of the 13 endnotes in the report, two are *op cits* to the Vermont and Kentucky studies. One source is from an independent publisher and another article is from the Hoover Institute's *Education Next* magazine. The remaining endnotes are from press statements and opinion articles.

IV. REVIEW OF THE REPORT'S METHODOLOGIES

In form and language use, the reviewed report's structure is more akin to a political document than a research report.

There is no original exploration of issues or primary research. The truncated review of other studies is deficient and does not qualify as a review or summary of the literature. Of the two studies examined, the report's presentation of one over-relies on a secondary analysis from another free-market think tank. In the other, the source is a brief summary of a larger effort. Particularly troublesome is the use of a select two studies, apparently chosen because they support the author's perspective.

The Tenuous String

Rather than a logical and inclusive examination of a key issue, the report's sections follow each other in a tenuous string. For example, in the introduction the author asserts that the House education chair "backpedaled" on portfolios. Even assuming the author's interpretation is correct, such a beginning neither frames the paper nor advances the author's contentions.

At the end of this initial section, the author presents the primary shortcoming of portfolio assessments as their lack of reliability (which term he carelessly interchanges with *validity*) from one rater to another. Yet, the quoted transcript of the chairman – the "evidence" on which this conclusion is purportedly based – is irrelevant to validity or reliability.

The next section, "The NEA's Push for Portfolios," may or may not be a true characterization.[14] The report presents a summary conclusion regarding NEA motives, which is not documented. The subsequent claim that the NEA is covering up students' failure to learn is likewise not supported by any evidence. An examination of the NEA's position paper on NCLB shows the organization as supportive of multiple indicators, but does not use the term portfolio.[15] As noted earlier, the phrase "performance or portfolio assessments" is used as a third tier descriptor in the NEA's legislative priorities, which is a separate document from its position paper on NCLB.[16] Such scant evidence provides a weak foundation for inferences about NEA intentions. Similarly, in the section titled "Concessions to the NEA," the key assertion is only the author's opinion that Representative Miller, "seemed to be yielding ground to the NEA" (p.3)

The next knot in the string is the puzzling tangent on workforce imperatives. The report's author attempts to link the NEA to the Commission on Skills and the American Workforce – the argument being that the Commission wants performance-based measures and the NEA wants portfolios. This interesting linkage is not substantiated and is remarkably off-point.

Attacking the Wrong Target

In statewide assessment programs and the NCLB reauthorization debates, the use of

portfolios for statewide high-stakes accountability purposes has received little attention. It is, therefore, puzzling why the Lexington Institute would attack such a minor target.

Methodological Issues Avoided

Both studies cited by Holland report low reliability coefficients for uniform statewide portfolio measurements and ascribe this drawback to vague prompts and lack of training and uniformity in scorers' decision-making. Such a finding breaks no new ground.

Researchers have studied and described a variety of issues and trade-offs associated with portfolio assessments. For instance, increasing traditional statistical reliability can be accomplished by any or all of the following: reducing the number of intervals in the rubric scale; training the scorers; redefining reliability to include the neighboring interval; scripting the assignment with such precision that it has a deleterious effect on the very qualitative concepts of interest. Exploration of these types of issues could have informed the examination of the potential use of portfolios as part of the NCLB accountability system. Holland and the Lexington Institute leave these matters unaddressed, however.

Portfolios as a Proxy for Multiple Indicators

Perhaps the report's most tenuous knot is the implicit equating of portfolio assessments with the "multiple indicators" clause under consideration in the NCLB reauthorization. These two terms cannot be substituted for each other. The characteristics are not interchangeable. In fact, each "multiple indicator" option will have its own profile of strengths and weaknesses.

The NEA position paper offers that multiple indicators

> could include . . . district-level assessments, graduation rates (for high school), attendance rates, school-level assessments performance portfolio assessments, and the percent of students participating in rigorous coursework, which may include dual enrollment, honors, AP or IB courses.[17]

While the NEA list presents portfolios as one of several exemplars, Chairman Miller's list does not use the term:

> Such additional indicators of school progress include graduation rates, dropout rates, percentages of students successfully completing end of course exams for college preparatory courses, assessments in history, science, civics and government, and writing, and improvements in the performance of the lowest and highest performing students in the school.[18]

Given this variety of approaches, the report's contentions regarding the shortcomings of portfolios (reliability, uneven implementation, differences in revision opportunities, differences in difficulty, cost, unevenness in teacher prompts, and variations in assistance and cheating) simply do not transfer to all of these non-test based indicators. For example, "percent of students in approved advance placement courses" is an indicator that is not prone to the criticisms associated with portfolios.

V. REVIEW OF THE VALIDITY OF THE FINDINGS AND CONCLUSIONS

Holland's strongest claims are that portfolio assessments for high-stakes accountability are unreliable and expensive. Yet, the published article by Koretz and his colleagues – which the Lexington author must be assumed to have read, since it's one of only two studies he cited – discusses two other studies that reported higher and more acceptable reliabilities.[19] This selective use of research suggests the author either intentionally slanted the evidence or was unacceptably cursory in his analysis.

Other research, conducted or reported in the years following the cited 1994 studies also raises direct questions about the Lexington Institute's conclusions. In examining large-scale portfolio results in language arts, mathematics and science, Wolfe found that reliability could be increased with more portfolio scores and cleaner rubrics.[20] A meta-analysis by Jiang *et al* found that performance assessments could reach acceptably reliability levels by care in the construction of the performance tasks.[21] In dealing directly with the cost/reliability problem, Parkes concluded that it is possible to develop reliable and cost-effective performance assessment systems.[22]

Accordingly, Holland's basic conclusions about portfolios are placed in doubt. And to then generalize to all multiple measures from this questionable base completely discredits his work.

Validity, Reliability and Costs

Unexamined in the Lexington report are serious concerns about the validity, reliability and considerable cost of the existing NCLB accountability system.

The current system is primarily driven by standardized tests.[23] These do not measure all the purposes of education, however. By concentrating on multiple choice tests in reading, math, and science, states cannot validly measure the full range of our purposes. Standardized tests are not valid indicators of these broader goals.[24]

In the author's school district, for example, a key measure of success for the experiential education program for troubled adolescents might be whether the students avoid pregnancy or arrest. In a less dramatic fashion, civic involvement of students could be a measure of success in the real-life of schools and communities.

Portfolios are an excellent tool for teachers in formative assessment, although they do not as readily lend themselves to comparing one school to another. Their greatest power is instructional improvement rather than summative evaluation.[25]

Traditional standardized tests can easily be made to show high statistical reliability by simple expedients such as increasing the number of items. When used in high-stakes applications, however, the reliability of the resulting decisions plummets. Kane and Staiger demonstrated that the error can exceed 70% of the variance in the existing system.[26] That is, each test has an error term, and when two tests are compared, the error terms are multiplied with each other. Furthermore, when this year's fourth graders are compared with the next year's fourth-graders, it is not clear whether the differences are due to the school or the differences between the two groups of students. The overall result is that the current NCLB system is extremely inaccurate.[27]

Holland correctly observes that portfolio assessments are costly. Palaich, however,

found that NCLB assessment and accountability systems cost states between two and three times what they receive in new federal funds.[28]

An honest concern about validity, reliability and costs must begin with an assessment of these same concerns for the system the author espouses.

VI. USEFULNESS OF THE REPORT FOR GUIDANCE OF POLICY AND PRACTICE

The report offers little useful guidance for policy or practice.

One core deficiency is that it examines an issue (portfolio assessments) which is not at or near the center of the debate. Multiple indicators are at the center. The reported shortcomings of portfolios are simply not generalizable to the other listed multiple indicators. In fact, the report's conclusions on the shortcomings of portfolios are open to question. These factors alone marginalize the utility of the report.

In practical terms, a state would first have to advance the notion of statewide portfolio assessment, overcome the technical and cost obstacles summarized by Holland, and then obtain federal approval to use the mechanism. While field trials in these areas would be a welcome expansion and would update our knowledge, the time and expense of such a process makes the likelihood remote.

For these reasons, using widespread use of portfolios in a high-stakes environment is receiving limited attention by state and federal policymakers.

While the Lexington study asserts that portfolio assessment is a step backward in school accountability, policy makers might well conclude that the broader exploration of multiple measures holds the promise of providing more valid and inclusive indicators of the entire spectrum of educational goals. For the furtherance of educational quality in a democracy, this would be a step forward.

NOTES & REFERENCES

[1] Holland, R. (2007, Sept. 17). "Portfolio Assessment: How to Weaken Accountability for $23 Billion in NCLB Spending." Lexington Institute. Retrieved Sept. 17, 2007, from http://lexingtoninstitute.org/1169.shtml

[2] H. R. ___, to reauthorize the Elementary and Secondary Education Act of 1965. Summary of Discussion Draft Retrieved Sept. 3, 2007 from http://edworkforce.house.gov/bills/MillerMcKeonNCLBDiscussionDraftSummary.pdf

[3] McMurrer, Jenifer (2007). *Choices, Changes, and Challenges: Curriculum and Instruction in the NCLB Era*. Washington: Center on Education Policy. Retrieved Sept. 3, 2007, from http://www.cep-dc.org/index.cfm?fuseaction=document.showDocumentByID&nodeID=1&DocumentID=212

[4] Nichols, Sharon L. & Berliner, David C. (2007). *Collateral Damage: How High-Stakes Testing Corrupts America's Schools*. Cambridge, MA: Harvard Education Press.

[5] National Education Association (2007, April 18). *Joint organizational statement on "No Child Left Behind."* Retrieved Sept. 3, 2007, from http://www.nea.org/presscenter/nclbjointstatement.html

[6] See, for example, the Aspen Institute's report: *Beyond NCLB: Fulfilling the Promise to Our Nation's Children*. (2007). Retrieved Sept. 3, 2007, from http://www.aspeninstitute.org/atf/cf/%7BDEB6F227-659B-4EC8-8F84-8DF23CA704F5%7D/NCLB_Book.pdf

[7] Holland, R. (2007) *Portfolios: a backward step in school accountability*. Arlington,VA: Lexington Institute. p. 1. Retrieved Sept. 3, 2007, from http://lexingtoninstitute.org/docs/holland_portfolio_assessment_8_29_07.pdf

[8] NEA's top legislative priorities for ESEA. March 21, 2007. Retrieved Sept. 3, 2007 from http://www.nea.org/esea/legpriorities.html

[9] Koretz, D., Stecher, B., Klein, S, and McCaffrey, D. (1994) The Vermont portfolio assessment program. *Educational Measurement: Issues and Practice*, *13*(3), pp. 5-16.

[10] Hambleton, R.K., Jaeger, R. M., Koretz, D., Linn, R.L., Millman, J., & Phillips, S. E. (1995, June 20). *Review of the measurement quality of the Kentucky instructional results information system, 1991-1994*. Office of Educational Accountability, Kentucky General Assembly. Retrieved Sept. 3, 2007 from http://www.lrc.ky.gov/oea/reports/MEASUREMENT%20QUALITY%20FINAL%20REPORT%2091-94.pdf

[11] Koretz, D., Stecher, B., Klein, S, and McCaffrey, D. (1994) The Vermont portfolio assessment program. *Educational Measurement: Issues and Practice*, *13*(3), pp. 5-16..

[12] The institute's mission is to ". . . champion freedom, opportunity, and personal responsibility for all individuals by advancing free market policy solutions." Retrieved Sept. 7, 2007, from http://liberty.pacificresearch.org/about/default.asp

[13] Koretz, D., Stecher, B., Klein, S, and McCaffrey, D. (1994) The Vermont portfolio assessment program. *Educational Measurement: Issues and Practice*, *13*(3), p. 11.

[14] *Note from the publishers*. This project receives indirect funding from the NEA. However, neither the NEA nor any of its representatives played any role in the selection, writing or editing of this review, and Dr. Mathis was not in a position to characterize NEA motives.

[15] National Education Association (undated) "NEA position on the 'No Child Left Behind' Act/ ESEA." Retrieved Sept. 3, 2007 from http://www.nea.org/esea/policy.html

[16] National Education Association (2007, March 21). "NEA's Top Legislative Priorities for ESEA," Retrieved Sept. 3, 2007 from http://www.nea.org/esea/legpriorities.html

[17] National Education Association (2007, March 21). "NEA's Top Legislative Priorities for ESEA," Retrieved Sept. 3, 2007 from http://www.nea.org/esea/legpriorities.html

[18] H. R. ___, to reauthorize the Elementary and Secondary Education Act of 1965. Summary of Discussion Draft Retrieved Sept. 3, 2007 from http://edworkforce.house.gov/bills/MillerMcKeonNCLBDiscussionDraftSummary.pdf

[19] Koretz, D., Stecher, B., Klein, S, and McCaffrey, D. (1994) The Vermont portfolio assessment program. *Educational Measurement: Issues and Practice*, *13*(3), p.11.

[20] Wolfe, E. G. (1996). *A report on the reliability of a large-scale portfolio assessment for language arts mathematics and science*. Paper presented at National council on measurement in education, New York. Retrieved Sept. 9, 2007 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/14/b0/f9.pdf

21 Jiang, Y.H. (1997). *Error sources influencing performance assessment reliability or generalizability: a meta analysis*. Paper presented at American educational research association, Chicago. Retrieved Sept. 9, 2007 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/b8/9e.pdf

[22] Parkes, J. (2000). The relationship between the reliability and cost of performance assessments. *Education policy analysis archives*, *8*(16). Retrieved Sept. 9, 2007 from http://epaa.asu.edu/epaa/v8n16/

[23] NCLB also requires that all students must be tested in science at least once within the grade spans K-5, 6-9 and 10-12.

[24] Rothstein, R. & Jacobson, R. (2006). The goals of education. *Phi Delta Kappan*. 88(4). December, pp 264-272.

[25] See, for example, "Portfolio assessment." Retrieved Sept. 9, 2007 from http://www.eduplace.com/rdg/res/literacy/assess6.html

See also, Koretz, D., Stecher, B., Klein, S, and McCaffrey, D. (1994) The Vermont portfolio assessment program. *Educational Measurement: Issues and Practice*, *13*(3), p. 12

[26] Kane, T. J., Staiger, D. O. & Geppert,J. (2002, spring). Randomly accountable. *Education Next*. Pp 57-61. Retrieved Sept. 3, 2007 from http://www.dartmouth.edu/~dstaiger/Papers/KaneStaigerGeppert_ednext2002.pdf

[27] To partially address this problem, proposals are now being considered whereby the same group of students would be tested as fourth graders and a year later as fifth graders. This "growth model" is logically sensible yet raises new problems with test equating. Even with growth models, which will likely be included in the reauthorized NCLB, the error rate is higher than 50%.

[28] Palaich, R. (2005, December 1). *Estimating the NCLB costs for states and school districts*. Paper presented to Communities for Quality Education, San Diego, CA.