HIGH STAKES TESTING IN K-12 SCHOOLS
AN ANNOTATED BIBLIOGRAPHY

TEST DEVELOPMENT AND ADMINISTRATION FOCUS

Introduction

This bibliography begins an ongoing effort to develop a resource for those interested in research on high stakes testing in K-12 schools. The bibliography is a work in progress, is not exhaustive, focuses primarily on empirical research, includes mostly references in the past decade, and includes multiple perspectives on the issues.

Feedback or suggestions for other entries (please send complete citation) for this bibliography should be sent to Sandra Mathison at smathison@louisville.edu.

Note: A number of people have contributed substantially to the preparation of this bibliography. They include Sandra Mathison, University of Louisville; Melissa Freeman, Kristen Wilcox, Lynee Sauer, University at Albany, SUNY. Preparation of this publication was supported under Grant # ESI-9911868 from the National Science Foundation. The contents do not necessarily reflect the position or policies of NSF.

Bibliography

Camilli, G., & Bulkley, K. (2001, March 4). Critique of "An Evaluation of the Florida A-Plus Accountability and School Choice Program." *Education Policy Analysis Archives, 9*(7). http://epaa.asu.edu/epaa/v9n7/

The Florida A-Plus accountability system uses scores from the Florida Comprehensive Assessment Test (FCAT) and student referral and dropout rates to assign schools one of 5 grades (A, B, C, D, F). An earlier evaluation of the accountability system, *An Evaluation of the Florida A-Plus Accountability and School Choice Program,* reported a high correlation between the threat of school vouchers and improved test scores. This critique takes a second look at that evaluation and suggests this correlation may be due to other factors such as sample selection, regression to the mean, how gain scores were combined across grade levels or how schools were used as units of analysis. These authors conclude that the evidence provided in the evaluation cannot support the conclusions that school vouchers are responsible for higher scores.

- Category: Student Achievement/ Special Populations; Test Development and Administration
- Keywords: Florida; FCAT; accountability; test results; validity

Cizek, G. (1996). Setting passing scores. *Educational Measurement: Issues and Practice, 15*(2), 20 – 31.

Cizek offers a procedural definition of standard setting focusing on the process of rationally deriving, consistently applying and describing procedures on which judgments can be made. Guidelines, models, methods, new modes of assessment and validity evidence in standard-setting are discussed. Test-centered, examinee-centered and compromise models are described. This article calls for measurement specialists to develop and refine procedures for setting standards on assessment.

- Category: Test Development and Administration
- Keywords: Validity; standards setting; assessment models

Coleman, A. L. (1998). Excellence and equity in education: High standards for high-stakes tests. [Electronic version]. *Virginia Journal of Social Policy & the Law, 6*(1), 81 – 114.

Discusses the latest wave of standards-based testing is that has moved from a measurement of minimum or basic skills to one of high standards learning for all. Coleman examines state educational reform efforts and issues regarding the fairness of testing practices as shaped by due process principles and anti-discrimination laws. Then he explores how the congruence or non-congruence between specific state standards, curriculum, instruction and tests affect the legal implications of educational decisions made based on such tests. He advocates for a more careful assessment of the design, administration and use of tests and their alignment with standards, curriculum, and instruction.

- Category: Historical/Political/Legal Contexts; Curriculum and Instruction; Test Development and Administration
- Keywords: Standards-based reform; policy decisions; civil rights; equity; discrimination; psychometrics

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. ***Educational and Psychological Measurement, 57*****(3), 373 – 399.**

This article examines pitfalls in the way new forms of high-stakes assessments use conventional analyses and interpretation of scores. Of concern is the use and misuse of appropriate measures of standard error or uncertainty of result. Student performances are part of larger measures of classrooms and schools. Therefore, the authors argue that concern over whether to treat individual scores as infinite measures or as measures limited to particular contexts needs careful consideration.

- Category: Test Development and Administration
- Keywords: Generalizability; reliability; standard error; score interpretation; student achievement

Guskey, T.R. (2001). High percentages are not the same as high standards. ***Phi Delta Kappan, 82*****(7), 534 - 536.**

Guskey asserts that setting cutoff percentages on tests is a complex process that goes beyond statistical formulas. He argues that cutoffs must be based on both teachers' judgments of the importance of the concepts addressed and consideration of the cognitive processing skills required to complete the task. Guskey makes the point that raising standards or increasing expectations for student learning cannot be accomplished by arbitrarily raising the cutoff percentages for performance levels or different grade categories. What is needed, he argues, is thoughtful examination of the tasks students are asked to complete and the questions they are asked to answer in order to demonstrate their learning.

- Category: Test Development and Administration
- Keywords: Standards; TAAS; cutoff percentage

Haertel, E. (1999). Validity arguments for high-stakes testing: in search of the evidence. ***Educational Measurement: Issues and Practice, 18*****(4), 5 - 9.**

This article is the Presidential Address given at the Annual Meeting of the National Council on Measurement in Education, Montreal, April 21, 1999. The author discusses three issues about validity in high-stakes testing. First, he provides an overview for how validity is currently determined. Second, he describes and presents a detailed validity argument for a large-scale testing program. Third, he suggests several strategies for studying and providing different perspectives as part of an ongoing evaluation of test validity. The purpose of listening to various stakeholders' assumptions about testing is to develop stronger validity arguments for high-stakes testing.

- Category: Test Development and Administration
- Keywords: Test validity; validity arguments

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. ***Educational Researcher, 20*****(5), 2 – 7.**

Until recently test scores were used for a limited set of purposes. The increased uses of standardized achievement scores for accountability purposes have increased opportunities for test score pollution. Test score pollution is based on contextual factors that alter test performance regardless of the construct the test intends to measure. Two major sources of test score pollution (student preparation and test administration

practices) are described. The authors believe that such pollution is pervasive in American education and suggest ways to combat test score pollution.

- Category: Test Development and Administration; Historical/Political/Legal Contexts
- Keywords: Test score pollution; accountability; cheating; test administration

Haney, W., Madaus, G., & Lyons, R. (1993*). The fractured marketplace for standardized testing*. Boston: Kluwer Academic Publishers.

The book scrutinizes the commercial aspect of testing and the effects of the marketplace on the quality of tests and test use. The authors find that monopoly markets prevail in some segments of the marketplace while in others small numbers of firms have oligopolistic control. The analysis ends with data from 1992, but the authors argue that the most relevant result of the book lies in the lesson that more care must be taken to avoid continuing to rely on imperfect test instruments arising from a highly fractured test market.

- Category: Test Development and Administration
- Keywords: Test validity; standardized testing industry; costs of testing; legislation; test preparation

Jacob, B. A., & Levitt, S. D. (2001). ***Rotten apples: An investigation of the prevalence and predictors of teacher cheating.*** **http://economics.uchicago.edu/download/teachercheat61.pdf**

As the stakes associated with standardized testing increases, there is a rising concern that teachers and administrators may feel compelled to cheat either in the way they administer the test or by altering student responses afterwards. In this study, Jacob and Levitt develop an algorithm for detecting teacher cheating on standardized tests. Using data from the Chicago Public Schools, the authors examine the question-by-question answers provided by students in grade 3 through 7 on the Iowa Test of Basic Skills (ITBS) for the years 1993-2000. Cheating is determined through two measures: unexpected test score fluctuations and unusual answer strings for students within a classroom. The authors found over 1,000 separate instances of classroom cheating which represents 4-5% of all classrooms in that district.

- Category: Test Development and Administration
- Keywords: cheating; ITBS; Chicago

Jenkins, J. A. (1993, April). ***Can quality program evaluation really take place in schools?*** **Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA. (ERIC Document Reproduction Service No. ED397067)**

Addressed in this paper is the accuracy of high stakes assessments. Jenkins explores whether high stakes assessment results are misleading due to extraneous factors. Some of the major problems with high stakes assessment he addresses are: lack of formal education regarding educational assessment; tests not being properly administered; possible lack of student motivation; students not prepared w/ test taking skills; and attention being paid only to scores. Jenkins includes changes that must be implemented before assessment reform will be successful.

- Category: Test Development and Administration
- Keywords: Test administration; measurement; student motivation

Johnson, E., Kimball, K., Olson Brown, S., & Anderson, D. (2001). A statewide review of the use of accommodations in large-scale, high-stakes assessments. *Exceptional Children, 67***(2), 251 - 264.**

Current standards-based reform efforts aim to educate all children including special education students. Advocates for students with disabilities support the inclusion of students with disabilities in state-mandated assessments and while most states provide accommodations for such inclusion, the psychometric, legal, and practical challenges of such inclusion are not well researched. This review finds that accommodations are provided but in ways that are inconsistent across districts and/or states. Specifically, the accommodations procedures used for the Washington Assessment for Student Learning (WASL) for 4th and 7th grade in 1998 are examined. These researchers conclude that the accommodations provided do not appear to place students in need of accommodations at an advantage over other students but do raise questions for discussion as to the ways various accommodations are determined and implemented.

- Category: Student Achievement/ Special Populations; Test Development and Administration
- Keywords: WASL; students with disabilities; test accommodations; test bias

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425 – 61.

The validity of test-based decisions is dependent on the appropriateness of the passing score and its relationship to the performance standard the test is designed to measure. Therefore validity is dependent on the interpretation of test scores not on the test item itself. This article outlines various approaches to the validation of performance standards, the setting of passing scores, and the assumptions upon which these are based. Kane then outlines the kind of evidence used to validate score interpretations providing the strengths and weaknesses of each.

- Category: Test Development and Administration
- Keywords: Performance standards; passing scores; score validation; arbitrariness of standards

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000, October 26). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8*(49). http://olam.ed.asu.edu/epaa/v8n49/

One of the reasons the high-stakes testing program in Texas, the Texas Assessment of Academic Skills (TAAS) has received so much attention is because of the reported large gains made by students. This study examines to what extent these reported scores are providing an accurate picture of student achievement in Texas. Comparison with student gains on the National Assessment of Educational Progress (NAEP) is provided as one way to assess the validity of Texas students' achievement gains. The authors conclude by raising serious questions about the validity of the gains in TAAS scores.

- Category: Test Development and Administration; Student Achievement/ Special Populations
- Keywords: Texas; TAAS; NAEP; gains; validity

Koretz, D. M. (1992). What happened to test scores, and why. *Educational Measurement: Issues and Practice, 11*(4), 7 – 11.

The author addresses two issues: "what has happened to the achievement of American students in recent decades, and what do we know about the causes of trends in scores?" Koretz discusses trends from World War II on, including the declines on achievement tests during the 1960s and 1970s and the upturn from roughly 1974 through 1980. The author investigates the trends' pervasiveness, timing or the "cohort effect", and differences among subject areas to find clues to understanding the broad patterns of trends in achievement on standardized tests. Koretz concludes that as the importance of test scores in the public debate continues to grow test data will increasingly be misused.

- Category: Test Development and Administration
- Keywords: SAT; Achievement trends; Minority students; NAEP

Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional testing on direct writing assessment. *Phi Delta Kappan, 80*(9), 673-679.

Mabry analyzes the contradictions between the direct assessment of student achievement in writing in classrooms and the state-mandated performance assessment. In particular, she contends that scoring rubrics are essential in large-scale and standards-based performance assessments in writing since they promote reliability assessments, but that the consequence is standardized writing as well. This in turn standardizes the teaching of writing and in the end the use of rubrics jeopardizes the teaching and learning of writing.

- Category: Curriculum and Instruction; Test Development and Administration
- Keywords: Rubrics; teaching to the test; pedagogy

Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education, 5*(3), 265 – 283.

As the stakes rise and unfavorable decisions are made about graduation, retention, employment and/or licensure based on a test, so does the possibility that a legal suit will be brought against test developers and users. This article examines such court cases along several dimensions to determine important and necessary standards for tests to withstand legal scrutiny. Several suggestions for test construction and use are offered.

- Category: Historical/Political/Legal Contexts; Test Development and Administration
- Keywords: Legal issues; validity; reliability; cut scores

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. ***Educational Measurement: Issues and Practice, 12*****(2), 9 – 15.**

The authors contend that one reason for the trend incorporating alternative assessment is partly due to current standardized testing practices—which measure achievement within too narrow a scope. This article explores the possibilities for introducing alternatives into state-wide curriculums. Examined in this article are (2) key components of using and interpreting alternative assessment practices:
1. psychometric properties (basis for use and interpretation)
2. consequences (reactions by student, teachers, administrators, those affected)
Also explained in this article are validity, fairness, cost, and reliability, as well as the consequences to alternative assessment.

- Category: Test Development and Administration
- Keywords: Alternative assessment; psychometrics; performance

Moore, W. P. (1994). The devaluation of standardized testing: One district's response to a mandated assessment. ***Applied Measurement in Education, 7,*** **343 – 367.**

This study examines teacher testing-related attitudes and practices in a court-ordered achievement testing setting. As part of a desegregation order, standardized tests were required as a measurement of effectiveness of the effort. The study provides background on the history of test use in the desegregation effort. The author explains, however, that desegregation is not assumed to be "*the* causative condition," but "contributes to the linkage between mandated reform and the need to show substantial improvement in student academic performance." The study showed that teachers were strongly influenced by the test to change their instructional efforts and curriculum, they were dissatisfied with the climate of pressure and accountability and overall, the results verify that teachers engage in inappropriate testing practices in order to improve test scores.

- Category: Test Development and Administration
- Keywords: Standardized tests; accountability; ITBS; teacher testing-related attitudes; desegregation

Moss, P. A. (1994). Validity in high stakes writing assessment: Problems and possibilities. ***Assessing Writing, 1*****(1), 109 - 128.**

Literacy education has had a strong influence on educational measurement and is largely responsible for the increased emphasis on writing in high-stakes assessments. However, there exists a tension between these two areas. In order to increase the validity of assessments used for high-stakes purposes the tasks conditions and scoring criteria need to be similar for all students. On the other hand, current literacy research suggests less standardized forms of assessment, those that favor more open and collaborative work between teachers and students are favored. A compromise that is often reached is to use less standardized forms of writing assessment in the classroom and reserve the more standardized forms for high-stakes use. Moss suggests that rather then make room for both in the curriculum the result is often a narrowing of the curricula as teachers prepare students to take the test. Moss offers several recommendations for assuring that good literacy and alternate assessment practices do not get buried beneath the pressure of high-stakes assessment practices.

- Category: Curriculum and Instruction; Test Development and Administration
- Keywords: Writing assessment; literacy practices; alternative assessment; portfolios

Paris, S., & McEvoy, A. (2000). Harmful and enduring effects of high-stakes testing. ***Issues in Education, 6*** **(1/2), 145 – 160.**

In this paper, the authors offer in-depth answers to reactions often voiced regarding high stakes testing. They also suggest methods for improving the current culture of high stakes—improvements that could be made at national, state, and local levels. Of primary concern is that testing procedures are being used and enforced as a way for policymakers to gain control over the educational system. It is the position of the authors that more research and monitoring be undertaken concerning the contexts of high-stakes testing.

- Category: Test Development and Administration
- Keywords: Accountability; high stakes reform; reaction/criticism; improving testing

Phelps, R. P. (1996). Are U.S. students the most heavily tested on earth? ***Educational Measurement: Issues and Practice, 15*****(3), 19-27.**

Both national data sources (the General Accounting Office and Organization for Economic Cooperation and Development) and international sources (the International Association for the Evaluation of Educational Achievement, the International Assessment of Educational Progress, and the Second International Mathematics and Science Study) provide mixed answers to questions regarding the character and extent of system-wide testing in the U. S. and thirteen, primarily European countries, permitting comparisons between nations. Findings indicate that U.S. tests tend to be in a multiple-choice format, of shorter duration, and placed at key transitions in the students' careers. U. S. students tend to take a greater number of individual administrations of short, norm-referenced system-wide tests with no or low stakes attached to them as compared to students in all thirteen other countries. U. S. students also take more classroom tests in mathematics and science than their international counterparts, but less in reading. Data from the aforementioned sources suggest that one could argue the point to the extent of testing in the U. S. from either side.

- Category: Test Development and Administration
- Keywords: SIMMS; system-wide testing; international comparisons

Shepard, L. A., & Dougherty, K. C. (1991). ***Effects of high-stakes testing on instruction.*** **(ERIC Document Reproduction Service No. ED337468)**

A brief overview is given of the evolution of standardized testing regarding student achievement—from standards being used solely as a tool for informing parents and monitoring trends to becoming one of the hottest debates surrounding educational reform. In this study teachers from two high-stakes districts were surveyed with questions concerning testing preparation and effects of testing on instruction. Some of the major findings of this study conclude: teachers feel pressure to improve test scores from administration and media; due to the emphasis on standardized tests, teachers focus on basic skills instruction; and four weeks of test preparation does not include the one to two weeks spent administrating the exams. Extensive tables are included in this study to illustrate its findings.

- Category: Curriculum and Instruction; Test Development and Administration
- Keywords: Teaching to the test; test preparation; instructional effects; pressure to improve scores

Smith, M. L., & Fey, P. (2000). Validity and accountability in high-stakes testing. ***Journal of Teacher Education, 51*****(5), 334 - 344.**

The authors of this article contend that in the current high stakes testing culture, accountability and validity are being placed in oppositional camps. Smith and Fey give very detailed definitions of each term. They also explain how accountability and validity under current educational reform plans effects teachers, assessment policies, and students. High stakes reform, according to them, produces instructional and assessment techniques that especially harm students who do not come from privileged backgrounds.

- Category: Test Development and Administration
- Keywords: ASAP; TAAS; Validity; accountability

Thompson, S. (2001). The authentic standards movement and its evil twin. ***Phi Delta Kappan, 82*****(5), 358 – 63.**

High-stakes testing reform is viewed as a medium for satisfying political agendas in this piece—agendas that

ultimately retract from learning and teaching experiences. Thompson compares and contrasts test-driven reform (high-stakes reform) to authentic, standards-based reform, and argues that no single exam should be used to evaluate a child's entire education.

- Category: Curriculum and Instruction; Test Development and Administration
- Keywords: Authentic reform; test-driven reform

Wongbundhit, Y. (1996). Administration of standardized competency tests: Does the testing environment make a difference? *ERS Spectrum, 14*(2), 3 – 8.

This article outlines optimal conditions for helping students perform well on standardized tests. Wongbundhit describes the administration of tests in Dade Country schools where the tests were given on Saturdays rather than during the school week in order to avoid overcrowding and distractions. The article presents a comparison of the school week administration in 1992 and the Saturday administration in 1993 in order to evaluate any differences in student performance, participation rates and test material security due to test administration factors. The study found that the Saturday administration provided optimal testing conditions with the effects of a positive impact on student performance; including in all subgroups. The study also found that the Saturday administration did not adversely affect the participation rates of students nor affect the security of test materials.

- Category: Test Development and Administration
- Keywords: HSCT; student performance; test participation rate; test material security; minority groups

Yeh, S. S. (2001). Tests worth teaching to: Constructing state-mandated tests that emphasize critical thinking. *Educational Researcher, 30*(9), 12 – 17.

Yeh argues that since test designers consider tests to lead instructional strategies, then it follows that developing test questions that foster critical thinking skills could push teaching in such a direction. He proposes conceptualizing critical thinking as 'careful argumentation' and suggests strategies for designing test items that push instruction in ways that would involve critical reading and discussion as preparation.

- Category: Test Development and Administration; Curriculum and Instruction
- Keywords: Critical thinking; test item construction; teaching to the test

Yen, W. M., & Ferrara, S. (1997) "The Maryland school performance assessment program: Performance assessment with psychometric quality suitable for high stakes usage". *Educational and Psychological Measurement, 57*(1), 60 – 84.

Implemented in 1991, the Maryland School Performance Assessment Program (MSPAP) is a performance-based testing program covering reading, writing, language usage, mathematics, science, and social studies. This study was conducted to ascertain whether the MSPAP has the psychometric characteristics necessary for use in high stakes decision-making. An outline of the scoring process, test difficulty, score validity, construct and consequential validity is provided as evidence that it does.

- Category: Test Development and Administration
- Keywords: Psychometrics; scaling; equating; score accuracy; validity

---