



PRESIDENT'S MESSAGE

Thel Kocher

The twin A's of accountability and assessment continue to garner unprecedented attention as a result of the No Child Left Behind Act. By the time you read this, the election will be over and we may or may not know who will be our president for the next four years. Regardless of the winner all indications are that NCLB will remain in place, perhaps with some modifications.

In this era of increased use of assessment results for high stakes decisions it is more important than ever that we as individuals, and through our membership in various organizations, make our voices heard regarding issues of fair and reasoned use of assessment results.

For several years your organization, NATD, has joined with AERA, APA, NCME, ACA, ASHA, and NASP as the Joint Committee on Testing Practices. The JCTP bylaws state that JCTP provides "a means by which professional organizations and test publishers can work together to improve the use of tests in assessment and appraisal." This is critical work and we can all be proud that our organization "has a seat at the table."

More recently, NATD supported the National Center for Education Outcomes at the University of Minnesota in their efforts to obtain funding to establish the Center for Research on Accessible Reading

Assessments. They were recently notified that funding was granted and NATD will be represented on their committees as they pursue this important work.

Certainly, our involvement in these two activities address several of the purposes of NATD. The NATD purposes are:

- ❖ To share information about testing in educational settings.
- ❖ To encourage the appropriate use of testing in educational settings.
- ❖ To improve the applications of measurement to students and educational programs.
- ❖ To encourage research in the area of elementary and secondary school testing and measurement.

Our annual meeting activities held in conjunction with AERA and NCME focus on these purposes, as well. This past April our speaker for the business meeting was Dr. Robert Linn, co-director of the Center for Research on Evaluation, Standards and Student Testing. Dr. Linn shared his views on "The No Child Left Behind Act of 2001: Accountability Requirements and Consequences."

Our symposium which is presented annually as an NCME session carried through with the AERA theme of Brown v. Board of Education: Fifty Years Later. Steve Schellenberg of the Saint Paul Public

Schools set the stage with his paper “Test Bias or Cultural Bias: Have We Really Learned Anything?” Stephen Sireci of the Center for Educational Research at the University of Massachusetts Amherst presented on “How Psychometricians Can Help Reduce the Achievement Gap: Or Can They?” Margaret Jorgensen of Harcourt Assessment, Inc. brought a test publisher perspective to the discussion with her paper “The Achievement Gap: Test Bias or Real Differences?” Jennifer McCreadie, then of the Indianapolis Public Schools, shared a district perspective through her paper. Glynn Ligon of ESP Solutions Group provided thoughtful comments in his role as discussant.

Joe O’Reilly of the Mesa Public Schools serves as our Symposium Proceedings editor and will be “publishing” the collected papers and discussant comments.

As you make your plans to attend AERA/NCME please consider arriving in time to attend the NATD Business Meeting which will be held on Monday morning prior to the official start of AERA. The Business meeting will be preceded by our annual breakfast compliments of NCS Pearson. Whether or not you can arrive in time to attend the business meeting, watch for the spring NATD newsletter which will carry information about the annual symposium. Typically the symposium is scheduled for a Wednesday time slot.

As we continue our journey toward the goals of NCLB, I encourage each of you to personally address the four purposes through activities such as:

- ❖ Presentations to community groups and school staffs.
- ❖ Service on committees at the state, regional or national level.

- ❖ Participation when local, state or national agencies or organizations solicit review and comment.
- ❖ Preparation of local guides to appropriate use of assessment results.



Thel Kocher, Director of Assessment & Evaluation, Edina Public Schools

NATD is a “mom and pop” organization. We have no paid staff; all work is carried out through the volunteer efforts of your officers and others. I also ask you to consider becoming more involved in NATD by volunteering to serve as an officer or by assisting in some other way. The NATD officers are listed below. Feel free to contact any one of us to discuss how you might assist the organization or to discuss items or issues that you would like for NATD to consider.

President—Thel Kocher, Edina Public Schools, thekocher@edina.k12.mn.us

President-elect—Peter Hendrickson, Evergreen School District, phendric@egreen.wednet.edu

Immediate Past President—Judy Levinson, Evanston Township High School, levinsonj@eths.k12.il.us

Treasurer—Sherry Rose-Bond, Columbus Public Schools, srosebond@columbus.k12.oh.us

Secretary—Michael Flicek, Natrona County School District, mike_flicek@ncsd.k12.wy.us

At-Large Board Members—

Bonnie Wilkerson, Northbrook School District 27, Wilkerson.b@nb27.org

George Olson, Appalachian State University, olsongh@appstate.edu

Phil Morse, Los Angeles Unified School District, phil.morse@lausd.net

Mike Strozeski, Richardson Independent School District, Mike.Strozeski@richardson.k12.tx.us

Co-Webmasters

Ray Fenton, Fenton Research, FentonResearch@aol.com

Phil Morse, Los Angeles Unified School District, phil.morse@lausd.net

Please plan to attend this session in Montreal!

NCME/NATD 2005 Symposium "Current Guidance for Integrity in Testing".

Presenters:

Karen E. Banks, Director of Research and Evaluation, Wake County Public School System, NC

"A Conceptual Framework for Judging Ethical Violations and Administering Sanctions"

This presentation will examine a conceptual framework for judging testing violations based on the North Carolina Testing Code of Ethics. A survey of current practices across several states will also be presented.

Gregory J. Cizek, Professor, University of North Carolina

"Personal and Systemic Influences on Integrity in Testing"

This presentation will focus on integrity in testing from two perspectives. First, integrity is often viewed from the perspective of personal actions that comport (or fail to comport) with ethical principles or established guidelines for testing. Second, though individual integrity is important, systemic conditions can also create an atmosphere in which personal integrity may be fostered or not. This second perspective--systemic influences--will be the explained and examples provided of current conditions that threaten integrity in testing. Finally, the presentation will examine psychometric practices that also serve--unwittingly--to attenuate the intended inferences that can be made based on test scores.

Jim Impara, Senior Director, Test Security Services, Caveon, Lincoln, NE and G. Gage Kingsbury, Director of Research, Northwest Evaluation Association, Portland, OR

"Cheating Detection within Computerized Adaptive Tests"

Operational adaptive tests create challenges for the detection of students who may be cheating and items that may be exposed. Since the tests are individualized and change dynamically, many common statistical approaches may not perform effectively. This study examines the performance of a state-of-the-art data forensics system in the identification of cheating on an adaptive test administered to middle school and high school students.

Discussant:

Joe O'Reilly, Director of Student Achievement Support, Mesa Public Schools, AZ

Organizer and Moderator:

Peter Hendrickson, Assessment Manager, Evergreen Public Schools, Vancouver, WA



High Stakes Test Preparation Materials Review White Paper Available

Contact: Peter Hendrickson, Ph.D., Assessment Manager, Evergreen Public Schools, PO Box 8910, Vancouver, WA, 98661 360 904 4015 phendric@egreen.wednet.edu

Washington Educational Research Assn. (WERA) and the Oregon Program Evaluators Network (OPEN) have jointly published a White Paper to assist test directors and others sorting through the many products and services that purport to help raise test scores on high stakes state tests.

Guidelines for Reviewing Test Preparation Materials is available online at www.wera-web.org as a .pdf document with references, a selected annotated bibliography and evaluation checklist for reviewing test preparation materials and services. The checklist was adapted from the Scriven (2000) Product Evaluation Checklist, a suggestion from the program evaluators.

Editor Gordon B. Ensign, Jr., noted, “Teachers and administrators are feeling increased pressure from state and federal accountability systems to improve students’ scores on tests. In response to this pressure, they often turn to special, sometimes questionable strategies and materials to help students prepare for tests. Many teachers and administrators are not clear about the line between proper and improper test preparation.” The paper is intended to provide guidance on those issues to the professionals in schools and assessment offices for “choosing materials as they best know the local educational context in which a particular set of strategies will be used.”

Ensign emphasized that “strategies for teaching to the test or test preparation materials are inappropriate if they raise test scores without also increasing students’ knowledge and skills in the broader subject domain being tested.” The paper also states that, “Federal, state and local policy makers (especially local school directors), that establish accountability requirements that focus on higher test scores rather than on improved student learning” must share responsibility for inappropriate test preparation practices.

The checklist asks raters to consider need, field trials/generalizability, efficacy, long-term effects, systematic evaluation, causation established, cost, and defensibility. Each area carries a zero to four rating with several points to consider.

Copies of the White Paper will be available at the Annual NATD/NCME Symposium, Current Guidance on Integrity in Testing, in Montreal. Committee members were Peter Hendrickson (Chair, Evergreen (WA) Public Schools), Tanya Ostrogorsky (Portland State University), Tom Owen (Consultant), Michael Ponder (Consultant), and Michael Power (Mercer Island (WA) School District).

Editor's Note. Each year many NATD members present papers at the annual AERA/NCME meetings. Summaries of two papers from last year's meetings in San Diego are featured below as examples of the contribution of our membership to these important meetings.

Meshing Assessment and Curriculum: A District Perspective

Duncan MacQuarrie

Jim Popham was invited to organize a symposium on “Meshing Measurement with Curriculum and Instruction” for the 2004 meeting of the National Council on Measurement in Education (NCME). He asked a representative from each of three settings, higher education institutions, state departments of education, and school districts, to proposed strategies reflective of their venue. I was asked to join the panel to provide the perspective of a district test director.

In districts where there are curriculum, instruction, and measurement specialists what strategies can test directors engage to enhance the likelihood that productive meshing will occur? Jim asked his panel to address this question in two ways. First, describe one or more ways of engendering stronger collaboration among educational professionals working in the fields of assessment, curriculum, and instruction. And second, identify some of the assessment content, concepts, skills, etc. that should be transmitted to instructional and curriculum specialists.

My approach to engendering stronger collaboration was built on honoring and respecting the work my curriculum and instruction colleagues do. That is, collaboration must be based on: having mutual respect for the value of the other's discipline – accepting that each specialist knows their area better and believes it is very important; learning as much as you can about the field of

curriculum and instruction – having an understanding of their world view; and a willingness to meet curriculum and instruction professionals on their turf – in the venues in which they operate.

You certainly need to build some “collaborative capital.” This can be a slow process, but some of the approaches that were successful for me were: being responsive, readily available when asked for help; being a patient and careful listener in order to find out what the real issues were, what kind of help or information was really being ask for and then making sure I followed through. You need to be willing to do almost anything in order to earn some “collaborative capital.” Remember, the curriculum and instruction folks are most likely busier than you are – in the early stages don't offer solutions that are likely to require them to do a lot of work, give them turn-key solutions whenever possible. Make sure you do as much of the “dirty work” of the testing program as possible, go the extra-mile to ease the burden of the assessment program. Do your job at the highest level so that curriculum and instruction professionals don't feel the need to clean up after you. And finally be aggressive in seeking opportunities to work together as a team with test interpretation, staff development, data reviews, etc.

So, what about the assessment content, concepts, and skill that we should help our colleagues in curriculum and

instruction understand and master? We should always be alert to the “teachable moments” when we can remind our colleagues of some of the big ideas in our discipline. Among many, I’m always looking for timely opportunities to remind my colleagues about some of the following concepts.

- a) Measurement is about quantifying mental constructs that we cannot directly observe (knowledge, understandings, thinking and reasoning processes, dispositions) and we do that by making inferences from behaviors we can observe. The tricks of our trade, our tests or assessments, are designed to provoke observable behaviors that provide good indicators or proxies for the constructs we can’t see. It is all about reasonable inferences and understanding the potential threats to the validity of those inferences.
- b) Even in the current climate of standards based measurement and the associated emphasis on classification of students, the ability to validly describe and documenting individual differences remains a very important component of good measurement.
- c) Reliability or consistency is an important, but not a sufficient condition for good measurement.
- d) Validity is most important, and validity is about the inferences we want to make from our measures. Therefore, our tests typically have multiple validities because our measures, test scores, are often used to support different decisions or inferences.

- e) The questions that make up our tests almost always reflect a sample of some much larger domain of interest.
- f) Tests, or more particularly improved test scores, should not be the target of instruction. The real targets must be the curriculum content that the tests are designed to reflect. This principle is based on the idea that our test scores are really proxies for things we cannot directly observe and that our tests only samples from a larger domain.
- g) Multiple measures and trends are fundamental to increasing the validity of our decisions.

There are also some important practical skills and knowledge we can help our colleagues develop and appreciate.

- a) Understand and expect that test development should be driven by test blueprints and item specification.
 - b) That there is an art and science to test and item development and the process is more complicated than it seems. They should appreciate that not everyone can be a really good item writer, but it is important to know that clear guidelines exist and they should know where to find them.
 - c) The basic principles for the design of valid scoring systems for constructed response items.
 - d) Some idea of what a comprehensive, but reasonable, classroom assessment model might look like.
 - e) What an alignment study is and the basics of how to design and conduct one – that is, how to
-

make sure that the test content and demands match up with the curriculum content and cognitive complexity.

As a final thought, I believe our curriculum and instruction colleagues should expect a lot more from measurement professionals when it comes to reporting and portraying the

results from our assessments, and they should demand that we do better work in this area.

Duncan MacQuarrie is the former Manager of Student Assessment for the Tacoma (WA) Public Schools and current is a National Measurement Consultant focusing on state accounts for Harcourt Assessment, Inc. He lives and works from his home in Olympia, WA.



So ... What's the Question?

Michael Flicek

School performance on large scale assessments can be summarized in a variety of ways. Regardless of the approach used to represent school performance on an assessment, it is usually possible to rank schools from high to low based on the results of the approach. Policy makers and the public are led to the conclusion that some schools are “better”, i.e., more effective, than others as a result of the reported summary. In truth, however, the conclusion about the school is often a function of the approach used. A school, for example, that ranks low with one approach may rank much higher with another approach. Furthermore, different approaches actually answer

different questions about schools. Such subtleties are typically lost upon lay consumers of the reports. In an effort to increase assessment literacy of the public and policy makers, I have found the use of a simple achievement-by-design matrix quite useful. Dale Carlson, an assessment consultant who served as the state assessment director for California for 17 years, was instrumental in initial work on this matrix (Carlson, 2002). Within each of the 4 quadrants of the matrix, assessment data are being used to answer different questions about schools. The achievement-by-design matrix is presented in Table 1.



Table 1.

Approaches to Measuring School Quality with Large Scale Assessment Results

	Achievement	
Design	Effectiveness	Improvement
Status	Quadrant 1 (Q1)	Quadrant 2 (Q2)
	How proficient are students on a particular day?	Are recent students more proficient than previous students?
Longitudinal	Quadrant 3 (Q3)	Quadrant 4 (Q4)
	Are individual students becoming more proficient over time?	Are recent students increasing proficiency more than past students?

On the achievement dimension, it is important to distinguish between effectiveness and improvement (Linn, 2003). A highly effective school might be improving very little or at a much slower rate than an ineffective school. When a school is highly effective, that school would probably be quite conservative about changing practices. If a school is changing very little in practices, it may be unrealistic to expect that the school would be improving beyond its already high level of effectiveness. As such, it is possible for a consistently highly effective school to have small or even slightly negative improvement. Conversely, schools that are low on effectiveness might have a heightened sense of urgency to change. Those changes could well lead to high level of improvement, in which case, the school might still score just average on effectiveness. Finally, it would be difficult to argue against sanctions for a school that was consistently low on both effectiveness and improvement.

On the design dimension, the status design requires just one score for each student whereas the longitudinal design requires multiple scores over time for

individual students. As a result the concept of effectiveness is quite different with each of the two designs. Having high mean scores for students in a particular grade at a school (or a high percentage of proficient students) at one point in time is considered to be effectiveness for the status design. With the longitudinal design, test results from at least two, and preferably 3 or more (Singer & Willett, 2003), school years for each student are required to measure effectiveness. Regardless of the level of test scores at a school, effectiveness in the longitudinal design is a function of the extent that individual students are growing in proficiency over time. The extent that a cohort of students at a school is more proficient in grade 4 than they were as second grade students is an example of effectiveness in Q3. Finally, Q2 improvement refers to the extent that more recent cohorts have different Q1 effectiveness than past cohorts and Q4 improvement refers to the extent that Q3 effectiveness changes over time for different cohorts of students.

Each of the four approaches has the potential to provide useful information to school improvement teams. The

approaches offer answers to important but different questions about how schools are performing. The Q1 approach, for instance, is focused on summarizing the level of proficiency of students in a school at any point in time and, as such, provides no control for prior achievement level of students. School level differences may be due to differences in prior achievement rather than to differences in school effectiveness. Nevertheless, the Q1 approach is useful. If, for example, a school is performing at a much higher level than other schools with similar demographics and size then perhaps changes should proceed with caution lest the changes might lead to lowered performance over time. In contrast, the school that is performing much lower than other schools with similar demographics and size might be well justified in having a sense of urgency around making changes intended to increase performance. Conclusions like these gain strength when the relevant differences have persisted over time. The failure to address demographics in Q1 is widely viewed as unfair by educators due to the absence of control for prior achievement. High versus low poverty schools are capable of having similar and high Q1 results. The

resources required to get those results might be quite different, however.

Q3 approaches include those that are often referred to as “value-added”. Because Q3 approaches use scores from multiple school years for each student, the initial score for each student can serve as a pretest to control for the prior achievement level of students. Having this control for prior achievement increases the internal validity of the design (Hill & DePascale, 2003). Q3 inferences about school effectiveness are more valid than Q1 inferences as a result. Within Q3, schools where students are increasing their proficiency at a faster rate are considered to be more effective. To answer the question about the extent that conclusions about a school will differ based upon the approach that is used, reading scores on the Northwest Evaluation Association tests from successive grade 6 cohorts from 26 elementary schools in one district were used. The cohorts ranged in size from 1563 to 1845 students. The summary statistics for each school using each approach were computed. Correlation coefficients for the Q1 and Q2 school values with the Q3 and Q4 school values for the 26 schools were then computed. The results are presented in Table 2.

Table 2

Pearson Correlation Coefficients for $N = 26$ Schools Comparing Quadrant 1 and Quadrant 2 Values with Quadrant 3 and Quadrant 4 Values on a Test of Reading for Subsequent Cohorts of Grade 6 Students.

	Quadrant 1	Quadrant 2
Quadrant 3	0.26	0.19
Quadrant 4	0.42*	0.25

* $p < .05$.

It's apparent from Table 2 that effectiveness conclusions about schools in Q1 versus Q3 were quite different just as improvement conclusions about schools from Q2 versus Q4 differed. These findings are understandable considering that different questions are being asked and answered within these quadrants. Furthermore, even though all students included in the analyses come from the same cohorts, different test data was needed to perform the analyses. Only current scores for individual students were needed for the Q1 and Q2 analyses while current scores and scores from the previous 2 years for individual students were required to perform the Q3 and Q4 analyses reported in Table 2.

The low cross quadrant correlations would be problematic if the purpose was high stakes accountability. When the purpose is to inform school improvement these low correlations are less problematic. Because different questions are being asked and answered within each quadrant, using analyses from different quadrants better informs school improvement decisions. What's important is to assure that policy makers, the public, and, most especially, school improvement committees, understand which questions are being asked and answered by a report. The achievement-by-design matrix is quite useful for this purpose. Indeed, when the purpose is to inform school improvement efforts, schools actually benefit from answers to all of the questions posed in the achievement-by-design matrix. This gives schools a richer understanding of

their strengths, weaknesses, and trends so that they can make better informed school improvement decisions.

References

Carlson, D. (2002). The focus of state educational accountability systems: Four methods of judging school quality and progress. In W. J. Erpenbach, et al., *Incorporating multiple measures of student performance into state accountability systems—A compendium of resources* (pp. 285-297). Washington, DC: Council of Chief State School Officers.

Hill, R. K. & DePascale, C. A. (2003). Reliability of No Child Left Behind accountability designs. In *Educational Measurement: Issues and Practice*. 3, 12-20.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. In *Educational Researcher*, 32, 3-13.

Singer, J.D. & Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.

Michael Flicek is the Director of Assessment & Research for Natrona County Schools in Casper, Wyoming.

For more information or to obtain a more complete paper on this topic contact the author at mike_flicek@ncsd.k12.wy.us.

National Association of Test Directors Membership Application Form

US Department of Revenue Taxpayer ID# 222659646

Please type or print neatly.

Name: Dr. Mrs. Ms. Mr. _____

Title: _____

Organization: _____

Mailing Address: _____

City: _____

State: _____ Zip: _____

Phone: _____ FAX: _____

E-Mail: _____

Check here to request that your director information not be published on our NATD web site: _____

Membership Category: (please check one)

_____ Active Member: Responsible for educational testing programs in settings not primarily for profit.

_____ Emeritus Member: Active NATD member for at least five years and no longer employed on full time basis.

_____ Associate Member: Not directly responsible for testing programs and/or involved in test development primarily for profit.

Annual dues are \$20.00. Please make checks payable to "NATD".

Mail your check and completed application to:
Ms. Sherry Rose-Bond, NATD Treasurer
Columbus Public Schools
1091 King Avenue
Columbus, OH 43212
