

National Association of Test Directors

1993 Symposia

This is the ninth volume of the published symposia, papers, and surveys of the National Association of Test Directors (NATD). Its publication serves an essential mission of NATD - to promote discussion and debate on testing matters from both a theoretical and practical perspective. In the spirit of that mission, the views expressed in this volume are those of the authors and not NAID. The papers in this volume were presented at the April, 1993 meeting of the National Council on Measurement in Education (NCME) in Atlanta, Georgia.

The editors wish to express appreciation to the members of the Board of the National Council on Measurement in Education for their continued support of the National Association of Test Director efforts. Special thanks go to Candy Elliot of Mesa Public Schools for her assistance in producing this volume.

DEDICATION

This is the first volume of the NATD Symposia not edited or co-edited by Peter Wolmut. Peter was instrumental in co-producing the first NATD Proceedings in 1985. Since then he has lovingly (if this word can describe Peter's public persona) cajoled papers and discussants' comments and slaved to produce an annual volume of great interest to all NATD members. Peter's professionalism, dedication, and commitment of time and energy to this publication and NATD has left some big shoes to be filled. For all that he has done and for the legacy that Peter has left NATD, we dedicate this volume to Peter.

Authors and Editors

Judith Arter

Northwest Regional Educational Laboratory, 101 SW Main, Ste 500, Portland, OR 97204

Paul Brown

5726 East 54th Street, Indianapolis, IN 46226

Linda Carstens

San Diego City Schools, 4100 Nonnal St., San Diego, CA 92103

Joe Hansen

Colorado Springs Public Schools, 1115 N El Paso St., Colorado Springs, CO 80903

M. Kevin Matter

Cherry Creek Public Schools, 4700 S. Yosemite St., Englewood, CO 80111

Joe McDonald

Education Dept, Box 1938, Brown University, Providence, RI 02912

Joseph O'Reilly

Mesa Public Schools, 549 North Stapley Dr., Mesa, AZ 85203

Carole Perbnan

Chicago Public Schools, 1047 W. Albion, Chicago, IL 60626

Lauress Wise

Defense Manpower Center, 99 Pacific St., Ste 155A, Monterey, CA 93940

Peter Wobnut

5824 NE 22 Ave., PO Box 11426, Portland, OR 97211

Table of Contents

Symposium I

(Testing ... Testing ...) Do we know where we are going? Have we been here before?? The scoop from the P. O. O. P. P. 's

Introduction

M. Kevin Matter, 1993-94 NATD President

The Roots of School Testing Programs

Paul Brown, 1984-85 NAID President

Riding the Measurement Waves

Peter Wolmut, 1985-86 NATD President

Assessment in the Year 2001: the darkness and the light

Joe Hansen, 1992-93 NATD President

Symposium II

Objectifying the Subjective: Rubrics, Scoring Guides, and Other Ways of Knowing

Introduction

M. Kevin Matter

Quantifying Quality: Results of the NATD survey on scoring rubrics

Carole Perlman

From the Bottom Up: Rubrics developed by teachers

Linda Carstens

Scoring Rubrics for Performance Tests: Lessons learned from job assessment in the military

Laurens Wise

Designing Scoring Rubrics for Performance Assessments: The heart of the matter

Judith Arter

Discussion

Joe McDonald

Appendix I

Sample Tasks and Rubrics Submitted by NATD Members in Response to 1992 Survey

Appendix II

NATD Publications and Surveys: 1995-1992

Symposium I

(..Testing ... Testing ...) Do We Know Where We're Going? Have We Been Here Before?? The Scoop from the P.O.O.P.P.'s

M. Kevin Matter, Organizer
Cherry Creek (CO) Schools

Some educational innovations are cosmetic makeovers, of ideas presented a generation earlier, with few improvements to the original idea. Thus, to maximize the current focus on assessment, we need to reflect upon how school testing programs developed, recognize past successes and failures in assessment, and assertively help plan the future of measurement in public education.

This symposium presented a historical overview of the origin of schools' testing programs, their current status, and future directions from the perspectives of three former NAID presidents -thus, the scoop from the P.O.O.P.P.'s (Perpetual Orders of Past Presidents). Paul Brown describes the roots of school testing programs, while Peter Wolmut addresses several myths about current assessment "innovations" and Joe Hansen looks into his crystal ball to visualize the possibilities of measurement in the past decade.

THE "ROOTS" OF SCHOOL TESTING PROGRAMS

Paul F. Brown
Measurement/Evaluation Consultant

The purpose of this paper will be to describe my personal observations of the development of the "roots" of standardized group testing as found in the public school milieu and the relationship between testing and professional organization development, particularly the National Association of Test Directors (NATD). The changes in testing are reflected in changes in emphases in NATD.

In my neighborhood there is an ancient oak tree rumored to have been growing prior to the first human placing a footprint in the area. Its roots are deep, strong and must cover acres of ground. The limbs are as thick as many full grown neighboring trees and reach out across the streets, and many lawns. The oak has survived tornado-strength winds and real tornadoes recently. In many ways this giant tree is symbolic of the growth and development of the standardized testing phenomenon. Testing, as we know it today, began as an acorn during World War I (the BIG war) in 1917. I shall try to trace some of the growth of standardized measurement development as viewed through the eyes of one practitioner.

Our history is rather short-just 76 years-when compared in other areas such as education. You probably recall one of your measurement classes citing this history. To quote from Psychological Testing (pp. 11-13) by Anne Anastasi:

"Group testing, like the first Binet scale, was developed to meet a pressing, practical need. When the United States entered World War I in 1917 a committee was appointed by the American Psychological Association to consider ways in which psychology might assist in the conduct of the war. This committee, under the direction of Robert M. Yerkes, recognized the rapid classification of the million and a half recruits with respect to general intellectual level.... It was in this setting that the first group intelligence test was developed. ...the Army psychologists drew on all available test materials, and especially on an unpublished group intelligence test prepared by Arthur S. Otis, which he turned over to the Army. The tests finally developed by the Army psychologists have come to be known as the Army Alpha and the Army Beta. The former was designed for general routine testing; the latter was a non-language scale employed with recruits who were unable to take a test in English. ...Even prior to World War I psychologists had begun to recognize the need of special aptitudes to supplement the global intelligence tests. These developed particularly for use in vocational counseling of industrial and military personnel. Among the most widely used are tests of mechanical, clerical, musical and artistic aptitudes."

My earliest contact with school testing programs was at age four. My mother, like most mothers, considered her three children to belong in the genius category, at least. Thus she applied for my admission to kindergarten. Since I was not yet at the required admission age, I was referred to take a standardized test to determine my eligibility for admission. From my later experience and my somewhat limited memory of the testing, I believe I was administered the Binet. Evidently I passed the test or the fact that my mother was president of the PTA, I was admitted. Even these many years later, testing is often used to determine early placement eligibility.

As a school psychologist I did not have too much faith in the group testing process. Later, respect for standardized testing and its use in the improvement of instruction developed. The teachers in one system in which I was employed duly recorded the group test results in the students' cumulative records and forgot them. At this time the school system had no organized program for using the test results to improve instruction. This was to occur later. The school psychologists felt that their products were more productive and useful since the tests were individually administered and teachers were provided with interpretation of the results and recommendations for their use.

At this time we were approaching the era of strong protestations against group testing in the schools primarily because of a concern that tests were not perceived as having eliminated bias. They felt that prejudicial results entered in students' records could cause excessive damage to individuals, particularly minority students. I have a feeling that my predecessor wisely chose this time to change to another branch of administration.

During this period an organization was developed which was a very helpful support group for test directors from the eighty largest school systems. The name of the organization evolved into Large School Systems Invitational Conference on Measurement in Education (LISSICOMIE). Possibly a little whimsy of Phil Clark, an organizer, entered into the name of the group. As a new test director, this group was an answer to prayers. It helped like no other organization at the time, to establish, a network of friends who could prevent each of us from reinventing the wheel. More importantly, we learned of new developments in the testing field and how other cities were dealing with the innovations.

The first meeting of LISSICOMIE consisted of hearing a very strongly resentful parent blame us and testing, in no uncertain terms, for all the evils in her school system. It was our baptism of fire in the field of protest. The era lasted for a long time.

LISSICOMEE was also to become one of the strong "roots" of what was to become the National Association of Test Directors (NATD). Prior to LISSICOMIE there was an informal group called Large City Test Directors (LCTD), all male, which met for dinner during the AREA/NCME annual meetings. According to unpublished notes by Tommy Hall, former test director of the Houston (TX) Independent School District, a group involved in testing had met at AASA between 1958 and 1962. After that the association was with AREA/NCME.

Testing, our "tree," was changing at least its upper structure from an acceptable branch into one which people were attempting to prune. During this period of protest the teachers' union was urging to delay, thus cancel, our testing program. Even the National Association of Elementary School Principals entered into the movement. In our school system not only was there no cancellation of the testing program but rather, at the urging of a Board member, the program was enlarged to include testing all students in kindergarten through grade twelve each year. Perhaps this was an inkling of what was to come in the accountability movement involving testing.

In 1972 the theme of the LISSICOMIE meeting was "Testing and communication in the urban setting: Problems, practices, and priorities." In our school system it was decided that this would be the year to release school by school test results. Another sip of accountability? The school system was the only one in the state to release any test results at all - a trend which continued until state-wide mandated testing was instituted and the state released all the results. I had learned a good lesson at LISSICOMIE and was thus better prepared and more knowledgeable about what to expect

We provided the media with a great deal of information prior to releasing test results. The print media were generally more thorough in their reporting and printed the entire press release covering at least three full pages of statistical data. I also prepared our principals and teachers for this event and encouraged the schools to also inform the officers of their PTOs prior to the onslaught of phone calls they would receive. The multiple pages of statistics evidently frightened even the most interested parents. Very few calls were received at the schools or the administrative center. We did have one interesting call from a student writer for a high school paper. She wanted the test publisher's phone number in order to verify that what we released was accurate. She is probably a reporter for 60 Minutes now!

One columnist spent weeks writing daily articles critical of our testing program and testing in general. His daily calls for further interpretations truly "made my day." Currently the media have become less stimulated by test data and thus coverage has decreased markedly. It seldom is front page news.

Colleagues in other cities which released test results informed us that the media reporting would move from front page news to a small item on the obituary page. How true!

In 1973 we studied our concerns about staff awareness of the role of measurement in the large school systems. A study revealed that many teachers and administrators received test results and promptly filed them for future reference. The next year mandated assessment was our concern.

Almost twenty years later we continue to be involved in mandated assessment but the focus has shifted from the school system to the state level and now possibly the federal level.

Later, in 1976, we faced the topic of standardized tests: "Criticisms within the Profession." This topic was to re-emerge in 1983 when a NATD committee studied large school system participation, or lack of it, in standardized test research. The results of this study were presented at AERA/NCME in Montreal, in April 1983.

Another development which created an impact on assessment was the arrival of criterion-referenced tests or a combination of criterion-referenced/norm-referenced tests merged into one instrument. Those school systems using this type of instrument needed to generate staff development programs to assist those administering tests in the process of being used as well as training in test interpretation for use in the schools and to inform parents effectively of the meaning of the results. Many individuals, both in the schools and in the community, felt they had an excellent understanding of norm-referenced tests and particularly grade equivalent and percentile scores and were having difficulty understanding the new tests. There was some skepticism felt, as indicated by comments of non-school individuals, that the use of criterion-referenced tests was one ploy used by educators to avoid accountability and to explain decreasing test scores in urban areas. The concepts of "mastery," non-mastery," and "partial mastery" were difficult to explain to many of the public. Many hours were spent explaining test results to parents who requested further interpretations of their child(ren)'s test results. As understanding grew so did acceptance. Most of the public still seemed to want to know precisely whether or not their children were below, at, or above the national norm rather than a listing of the objectives their children had mastered.

Concurrent with the arrival of criterion-referenced testing was the era of closer coordination of between the divisions of curriculum and the divisions of testing and evaluation. Test results provided some direction to the evaluation of student progress in attaining the objectives contained in the curricula of the grade levels. Some felt that testing was controlling curricula and that teaching was being directed to stressing and teaching only those objectives contained in the tests. Since testing was becoming more

related to accountability, one can understand the source of the analysis cited above. This feeling was reinforced when state-wide testing programs were being developed and installed. No longer were test results a school system product, but state-wide testing programs were being developed and installed. No longer were test results a school system product, but state-wide testing could lead to comparisons among school districts in the state. Real estate agents were beginning to use test results as a measure to attract the sale of homes in certain sections of the city or state. Parents with school age children moving into a new area could request to review the test results for all schools within an area with the focus of buying a home in a high achieving district.

Accountability, like discipline, is not a negative term. When accountability arrived on the education scene it was seen as a procedure which could be used to evaluate school systems, administrators, and teachers. This process has been used in business and industry as a method of improving productivity, sales and profits. It is a procedure which may be employed to improve instruction and student performance. It took some time to convince educators that accountability was a means of generating improved school functioning. During the era of public suspicion concerning the productivity of the nation's educational process, accountability became a viable technique for demonstrating to the public that educators were concerned about performance and were doing everything possible to improve. It became a less threatening process.

Tommy Hall's history of NATD indicates that NATD has been seriously involved in development in the measurement field as evidenced by the activities of practitioners in the nation's school systems. NATD members are responsible for the actual testing of millions of students each year, and thus are extremely aware of both the positive and the negative impacts of assessment on a first hand basis.

The organizations with whom NATD met for many years suggests that testing practitioners were seeking an organization or organizations with whom we could become allied which would be mutually nurturing. In 1984 Nancy Cole, NCME President and Barbara Plake, Program Chair invited NATD to offer a symposium at the annual meeting in April in New Orleans, Louisiana. George Madaus, President of NCME, greatly encouraged the growth of NATD in its relationships with NCME. We all are stronger because of the close relationship which developed among AERA, NCME and NATD. There is a strong symbiotic relationship among those who develop tests, conduct research in measurement, those who employ tests to improve instruction, and school systems which use the information.

Testing will continue to change. At times we will go back to procedures of the past. At times we will develop new techniques in measurement. Generally there will be an interaction of many branches involved in changes, developments and improvements.

References

Anastasi, Anne. (1968). *Psychological Testing*. New York: Macmillan Publishing Co., Inc..

Brown, Paul F, & Hall, J. (1983). "A Study of Large School System Participation in Standardized Test Research." Presented at AERA/NCME. Montreal, Canada, 1983.

Hall, Tommy. (1992.) Personal Communication, unpublished notes on LCTD, LISSICOME and NAID.

National Association of Test Directors. (1984). *NATD Newsletter*, .1 (2).

RIDING THE MEASUREMENT WAVES *

Peter Wolmut
Multnomah (OR) ESD

"After a few attempts at building objective tests to measure complex skills and understandings, the typical teacher is likely to return to the familiar essay examination whenever s/he wishes to do more than measure the student's knowledge of facts."

-William E. Coffman (1972)

Kevin invited me to share anything of import in our field which I may have discovered these past 32 years prior to my retirement at the end of this school year. The first thing I must admit is how humbling it is to realize how little one has to share! But my thoughts returned to a recent incident at one of the ECS Boulder conferences, when a group of us were quaffing our silver and gold bullets. I expressed my hope that the current wave of performance assessment "wouldn't go into the dumpster" like the last one did; and that I feared it would if its burden were again expected to be borne by teachers. One "young whippersnapper" (It's fun to say that just before retirement!) declared that I didn't know what I was talking about. Even though I should have recognized main bullet effects, I rejoined that I knew because I had lived through the last wave. The response was, "Who cares? This is a NEW performance assessment wave!" And so the topic I chose for today is the rhythmical sine-wave along which pedagogical method is promulgated, goes away, reappears under new nomenclature a few years later, is dropped, repromulgated, etc. Common to the promulgations is the declaration that this latest way of doing things will solve some nasty educational problem. The method is always presented as "NEW." And closely related discoveries from previous waves are left a-moldering while precious current research time is spent rediscovering many of the same things; or, worse yet, current practice is in error since ignorance of the previous wave(s) is bliss!

*The author wishes to express his deepest appreciation to Mary Bush MLS for her generous assistance in preparing this paper.

Continuing with performance assessment as the example, in an OERI performance assessment consumer guide (Sweet, 1992) the following is directed to the entire nation:

"Performance testing ... is a form of testing that requires students to perform a task rather than select an answer from a ready-made list. Experienced raters ... judge the quality of the student's work based on an agreed-upon set of criteria. This NEW [emphasis added] form of assessment is most widely used to directly assess writing ability based on text produced by students under test instructions."

One aspect of this document is that no references for its content are given. The reader is directed to a dozen persons to whom to write. These contemporaneous individuals are engaged in performance assessments.

Is Sweet alone in describing the wave as new? One has only to examine recent work in this field. To narrow the arena, I chose the topic of reliability. While indeed there are new studies dealing with reliability and IRT, a cursory and inexhaustive examination of AERA and NCME programs for this and the past three years shows the following titles:

"Cross-Scorer and Cross-Task Comparability of Judgments of Students' Reading, Writing, and Math Performance"

'Estimating the Reliability of a Direct Measure of Writing Using Generalizability Theory"

"Generalizability of a Statewide Performance Assessment"

"How Can We Ensure Accurate and Reliable Data from Authentic Assessment, or Can We?"

"Procedures for Establishing the Validity and Reliability of Performance Assessments"

"Reliability of Performance Assessment"

"Reliability, Validity, and Alternative Approaches to Assessment"

"Reliability, Validity, and Manageability in Large Scale Performance Assessment"

"Test-Retest Reliability in Holistically Evaluated Writing Samples"

Similar cursory review of the literature from the past two decades includes these matters and uncovers an interesting diamond or two. Coffman (1972) demonstrated multiple sources of error on essay tests just as in all measurement. In the process, he cited Stanley's (1962) ANOVA methodology which was one of many precursors to the famous generalizability theory of Cronbach et al (1972). That theory was

applied by Steele (1979) in a study of both holistic and analytical scoring to determine what happened when multiple samples and when multiple raters were used. And Rentz (1980) about then did one of his famous "let's discuss it in English" applications - this time of generalizability theory to reliability.

Tollefson & Tracy (1979) examined the scoring of well and poorly written social studies essays of varying lengths and found that long poorly written and short well written ones showed little difference in scores. In your essay system, are constraints placed on length to avoid duplicating these pejorative findings?

Developing a standardized writing test as part of the ITBS, Cantor & Hoover (1986) studied reliability, among other matters. Their gross finding was that rater reliability alone should not be used as an estimate of test reliability, in that lower reliabilities appeared both within and between modes of writing. The former has implications for making certain that students are instructed in a mode, while the latter undergirds the concept that a single mode should not be used as an estimate of total writing competency. This supported some of the findings of Michael et al. (1980), in which parallel questions within one mode focusing on different objects showed different results. Today there is an agency which assesses its students via multiple modes, whereafter the scores for each school are summed and divided by N! Might there possibly be another agency elsewhere following the same poor practice?

In the last review of research on writing, Huot (1990) asserted that the nature of the condition of much of writing literature was fragmented and ad hoc. However, he was confident of the work of the early statisticians, especially Paul Diederich in the sixties and seventies at ETS.

There is Nothing New about the Idea of Performance Assessment!

Over the past twenty years, research has answered many questions about good and bad practices. A mistaken notion of newness of a topic cannot be used to avoid reviewing older, as well as current, literature when performing a study. As part of the introduction to their paper, for example, Cantor & Hoover cited 21 published studies which dealt with essay test reliability and validity. Of these, ten were five years old or less; the other eleven dated from 1966 to 1981.

Through the use of a local University library or even with an ERIC CD-ROM system, preferably with the aid of a research librarian, it becomes relatively easy to track the history of the current wave and to take soundings every 5 years or so to see what that history reveals. To do less leaves us in an unenviable unethical quagmire.

References

- Cantor, N. R. & Hoover, H. D. (1986). The Reliability and Validity of Writing Assessment: An Investigation of Rater, Prompt Within Mode, and Prompt Between Mode Sources of Error. Paper presented at meeting of the American Educational Research Association, San Francisco, CA.
- Coffman, W. E. (1972). On the Reliability of Ratings of Essay Examinations. *Measurement in Education*, 1 (3).
- Cronbach, L. J., Gleser, G. C., Nanda, H., NaJaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley and Sons.
- Huot, B. (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research*, 0, 237-263.
- Michael, W. B., Cooper, T., Shaffer, P., & Wallis, E. (1980). A Comparison of the Reliability and Validity of Ratings of Student Performance on Essay Examinations by Professors of English and by Professors in Other Disciplines. *Educational and Psychological Measurement*, 4Q, 183195.
- Rentz, R. R. (1980). Rules of Thumb for Estimating Reliability Coefficients Using Generalizability Theory. *Educational and Psychological Measurement*, 41576-592.
- Stanley, J. C. (1962). Analysis-of-variance Principles Applied to the Grading of Essay Tests. *Journal of Experimental Education*, 32, 279-283. .
- Steele, I M. (1979). The Assessment of Writing Proficiency via Qualitative Ratings of Writing Samples. Paper presented at Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Sweet, D. (1992). Performance Assessment. OERI Educadon Research Consumer Guide No. 2.
- Tollefson, N. & Tracy, D. B. (1979). Response Length and Quality in the Grading of Essay Tests. Paper presented at annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Assessment in the Year 2001: The Darkness and the Light

Joe B. Hansen
Colorado Springs Public Schools

This paper is an attempt to create two alternative, and opposing visions of the future, in which the current movement toward performance based assessment has developed to a higher level of maturity and has effected change in educational practice and policy. These alternative views are not intended to foretell the future. Rather they are intended to increase the awareness of those who are willing to suspend their disbelief of these alternatives as possible consequences of this movement and by so doing, provoke thoughtful discourse on the direction, velocity and potential effects of the movement.

The idea for this approach came from a combination of two sources. The first was my experience as a participant and trainer in the "Consensus Process", a conflict resolution process in which people who are experiencing conflict are brought together to resolve it. This process requires that participants identify and describe their worst fears regarding the conflict situation. Having done this they are also instructed to articulate their best hopes for the situation. Working through the worst fears, enables them to think productively about how to attain their best hopes. It is an axiom of the Consensus Process that one must "go slow to go fast," meaning that when we rush into things without thoughtfulness and awareness of the pitfalls that might lie ahead, we are doomed to make mistakes that will slow our progress.

The second experience that led me to this approach came at the 1992 October Assessment Institute of the Northwest Evaluation Association, in Portland Oregon. NWEA Executive Director, Allan Olson asked me to serve on a panel where I was to create a vision of assessment in the year 1999. I found that my consensus training would not allow me to see a best-case vision without entertaining the possibility of the alternative worst-case vision. My hastily scribbled notes for that session contained the nucleus of ideas for this paper.

It is my best hope that these alternative scenarios will reveal the potential for both educational improvement and degradation that are inherent in this movement toward increased use of performance based assessment (PBA), for it has been said that every movement for reform contains the seeds of its own destruction. Whether this movement results in broad-scale or degradation will, as I hope to illustrate, depend on contextual factors as well as on the response of the professional educational community to this movement.

INTO THE DARKNESS: A Confession of My Worst Fears About Assessment in the Future

Underlying Context Assumptions

The future of educational assessment will depend on numerous political, economic and educational context factors. In order to create a vision of the future one must establish the context by examining those factors in advance. The following assumptions establish the context for the worst-case scenario of the future of assessment. It is now 2001.

President Clinton's economic recovery program was a dismal failure. With the stimulus component eviscerated by overzealous conservative elements who wanted deeper cuts, the plan bogged down in higher unemployment and lower than expected tax revenues. The resultant downward spiral of the economy dragged the economy into a steep recession led by high unemployment and a growing, rather than shrinking deficit. Anti-taxation sentiment has thus grown stronger, leading to even further reductions in support for educational funding and has re-fueled the fervor for a voucher system.

State and federal accountability requirements have increased while educational funding has continued its down-slide. All fifty states now have legislation mandating accountability systems based on performance assessments. Few, if any, provide funding support or positive incentives for school districts to participate in these programs, but virtually all of them bear sanctions for those districts that do not comply.

The continued fervor for educational reform has led to a proliferation of mandates by state legislatures and state boards of education for certified diplomas, certificates of mastery and the like, all undergirded by mandated systems of performance assessments.

Worst Fears for the Future of Assessment

School budgets, ravaged by vouchers, choice plans and tax limitation amendments have been expunged of research, assessment and evaluation funds. Among the consequences of the reduced funding for public education has been the elimination or severe reduction of the research, evaluation and testing departments of school districts. Assessment development, mandated by many state legislatures, but not funded, has become the responsibility of the curriculum specialists and staff development specialists, whose budgets have been spared so that they might develop and implement the mandated PBAs. As a result, the anti-scientism that characterized the test bashing of the late eighties has mostly supplanted the psychometric tradition. The staff development and curriculum specialists have led the charge, rejecting cries from the psychometric community for technical standards and plunging ahead with more "authentic"

but, unknown to them, less reliable and valid assessment tools. College and university teacher training programs, woefully under-funded themselves, have not developed the courses and still have no requirement for assessment training in their teacher certification programs. Very few teachers have been trained in the use of PBAs and those few who have received even the most minimal exposure to training are held responsible for training all others in a sort of "trickle down, share the ignorance" model.

NRTs and CRTs have become virtually extinct - as have the spotted owl, African elephant, great white shark, bottle nosed dolphin, and numerous species of birds and insects. Desertification has established itself where the Brazilian rain forests once stood. Due to the shrinking budgets and increased pressure for performance assessments, virtually all assessments are now performance assessment like, with little regard for traditional parameters of quality associated with classical measurement theory, e.g. reliability and validity; these terms having all but disappeared from the vocabulary of assessment except as subjects of derisive humor in workshops and many teachers' lounges.

The national standards and goals initiated under the Bush administration have been revived and have become institutionalized under the newly elected Quayle administration. This has resulted in:

- > A nationally mandated, every student national assessment program (NAEP) that is entirely performance based is funded by congress. Its results are reported at individual, school, district, and national levels. State legislators are using its results to beat up local boards of education and superintendents, thereby exacerbating the now monumental education crisis.
- > Governors are vying for bragging rights over their NAEP scores while congress is using NAEP data to allocate resources on a "reward the winners and punish the sinners" paradigm
- > All across the nation local boards of education, in response to public pressure and state mandates, have imposed performance standards and assessments on all schools in a top-down fashion. The result is a growing resentment toward local, state and federal mandates that has generalized to the method of assessment associated with the mandate. A counter-trend to bring back the more cost efficient and easily administered standardized, multiple choice tests has begun among educators and taxpaying citizens.

Meanwhile, back in the classroom:

Teachers' enthusiasm for PBAs, once ecstatic, has waned to a level of compliance with state and national mandates, in which they wearily conduct, in a ritualistic fashion, "cookie cutter" PBAs taken from

publishers' commercial banks of tasks and rubrics that have salvaged the fortunes of CTB-McGraw Hill, Riverside, The Psychological Corporation, ETS and others.

Resourceful students have taken advantage of low cost electronic media and interactive home television to "procure," produce and disseminate on a massive scale "Cliff Notes" versions of performance assessments keyed to "higher order thinking skills". A total breakdown of test security coupled with a lack of commonly accepted technical standards for PBAs has led to the development and rapid growth of an entire industry based on catalogues of PBAs that have succeeded in getting high school students "certified."

The cultural bias of PBAs has resulted in an increased disparity between those students from homes rich in stimuli such as books, magazines, family discussions of ideas and issues, etc. and those from less enriched environments, i.e. the less affluent. This has resulted in a decline in graduation rates, increased drop-out rates and has reinforced the myth that public schools are inferior, thereby further fueling the fervor for privatization of the educational system. The entire educational system is caught in a self perpetuating vortex of failure, criticism and withdrawal of funding.

Rick Stiggins, his dreams and hopes for educational assessment reform having been shattered, has retired in frustration and was last seen wading up an Alaskan stream, fly rod poised for a long cast under a far off grassy bank.

OUT OF THE DARKNESS AND INTO THE LIGHT

Underlying Context Assumptions

The following assumptions undergird the more positive or lightness view that characterizes the best-case scenario of the future of assessment.

President Clinton's economic plan was a smashing success. In its first term, the Clinton-Gore administration increased support to public education through expanded entitlements and increased competitive grants. Educational policy encouraged experimentation, supported innovations with funding and created an atmosphere of trust and support in which risk-taking could thrive. In its second term the Clinton-Gore administration continued the legacy with increased funding for educational research and applications of technology. As 2001 arrives the Gore administration has extended and strengthened this commitment.

> Accountability has taken on a longer term perspective, thereby reducing the pressure on the system for quick fixes. Federal and state governments have come to recognize that meaningful reform takes time, buy-in, collaboration and resources, and if it is to succeed, must be accompanied by cybernetic style data collection and reporting systems that provide continuous, valid and reliable feedback on a school district's progress in meeting locally determined standards.

> The Goals of America 2000, while still an active part of the national agenda, are not the driving force for reform that they were intended to be, but, because of the renewed emphasis on educational research funding and the application of research results in the classroom, significant progress has been made in reaching the goals.

> National standards have been established in math, reading, English, social studies, science and the arts. These standards have provided the framework for the development of assessment systems.

> The national focus has shifted from getting the "biggest bang for the buck" to an emphasis on quality education - serving the diverse needs of all students. Therefore:

Educational programs are more diverse and varied than ever before.

Cross-disciplinary, ungraded, continuous progress schools, connected with their communities through local citizen advisory boards (LCAB) are the norm. These LCABs are responsible for.

- analyzing the needs of the school
- recommending priorities to the principal and staff
- setting school goals and standards
- monitoring progress toward school goals and standards
- securing and screening volunteers from the community to work in the schools as tutors, teacher assistants, "grandfriends," grant writers, etc. advocating for the school.

These LCABS are supported by the district general fund budget at a modest level, for expenses only.

They are also eligible for state and federal incentive grants.

ASSESSMENT - A Best-Case Scenario

The attitude of politicians and other policy makers toward assessment has shifted toward increased emphasis on instructional improvement and a decreased emphasis on accountability. This is manifested in a reduction in state and federal mandates for accountability and an increase in incentives for improved outcomes, that are relevant to a student's future life and are supported by credible evidence. Incentives

vary in form, but include relaxation of restrictive regulations, cash awards for schools and positive publicity.

At the District Level

A systems based approach to assessment has begun to supplant the "one size fits all" mentality of the early performance assessment zealots. Increased funding support for educational research, evaluation and development has enabled district-wide testing programs to evolve from the old NRT survey achievement tests to comprehensive systems that include item-bank driven IRT based tests designed to measure progress on local curriculum goals, performance based assessments and portfolio assessments at classroom level and NRT programs are used on a sampling basis to provide policy makers with comparison data, Increased use of multi-dimensional (comprehensive) assessment systems also results in a much greater use of multiple types of data in decision making in order to increase validity and reduce risks that could lead to serious errors in decisions regarding the lives of students. An increased understanding of the limitations of certain types of performance assessments, based on experience and research findings has helped to reduce the over-reliance on performance assessments that began to replace the over-reliance on NRTs in the early nineties. Balance and moderation have begun to replace blind zeal and faddism. Electronic networks interconnect school district item banks across the country. A U.S.E.D. funded, research database on performance assessment is also accessible electronically. This database includes a resource bank of well researched "engaging tasks" and scoring rubrics that have been cross referenced to content and performance standards. This database is developmental and dynamic, continually receiving new assessments that have passed rigorous technical standards. The technical standards themselves have been developed cooperatively by the joint committee on evaluation standards and have been approved by ASCD, AASA, NEA and the AFT.

At The Classroom Level

Electronic technological breakthroughs, made possible by Clinton's Technology Initiative, have enabled teachers to collect, compile and use data much more efficiently. Their role is undergoing a transformation to a greater emphasis on being information managers and true learning facilitators. All teachers have access to the resource banks of assessments and scoring rubrics that have passed rigorous technical standards, established by cross-disciplinary standards boards.

At the classroom level technically sound performance assessments are integrated with instruction so completely that the words "test," "quiz," "mid-term" and "final" have all but disappeared from the educational lexicon. Assessment has become almost completely transparent to students. Teachers have learned to collect and use data in a continuous fashion rather than at specific points in time. This continuous stream of assessment data is facilitated by widespread use of CD-ROM and optical scanning technology, which is now becoming commonplace in America's schools.

Teacher Certification

Teacher preparation programs at universities are including a one-year sequence on educational testing and assessment in their required core program for teacher certification. This sequence includes psychometric theory as well as guidance in the development and use of performance assessments in the classroom. Knowledge of how to use a variety of assessment data appropriately in educational decision making is an exit outcome for teacher certification programs. Updating of this knowledge is a standard for certificate maintenance.

The Student's Perspective

"Quality" is a value precept that is instilled in students from their first day in school. Children learn to take pride in producing quality work and performing at a level of excellence that is appropriate for them. "Self concept is subordinated to quality performance. Students are taught from their entry into school what quality performance looks like and how it is judged, so that they can learn to monitor their own progress and strive for continuous improvement. Students begin learning from the day they enter school how to monitor their own progress by focusing on what they do well. Self reflection is integral to instruction. All students keep growth portfolios, which are increasingly becoming electronic. These growth portfolios are key components in student's educational growth portfolios, which contain a selective collection of evidence of each student's progress through the system. The evidence includes results of PBAs, CRTs, NRTs, educator's observations and student self reflections.

Peer review groups are used extensively and appropriately at all levels of schooling to help students learn to evaluate quality work and inculcate the value of excellence into their own value system.

National Assessment

National assessment based on NAEP has evolved into a balanced program of 1/3 standard NRT/ MC items, 1/3 open ended and performance type items, and 1/3 local assessments linked to a common

scale. It is now mandatory at state level, using a matrix sampling approach, optional district-wide on a local district level, at district expense. A trend has developed for states to link their state assessments to NAEP.

Societal Involvement

Employers are now reporting that high school graduates are entering the work force with the essential skills and core knowledge required to succeed as productive employees in an increasingly technologically sophisticated work place. Recent Department of Labor survey data reveals a significant decrease in dollars spent by employers on remedial education for employees. Meanwhile, the public has taken notice of these changes and have become staunch supporters of our public education system. The cry for a voucher system of education has been stifled. That vast majority of Americans without children in school turn out in great numbers to vote in support of educational funding measures.

All of these changes have resulted in an educational renaissance in which students are excited about school and teachers feel a new wave of joy and enthusiasm for their profession. The voting, tax paying public that only five or six years earlier was telling education to cut waste, reduce costs and improve student outcomes or face a voucher system has become engaged in the educational process at the local school level and is helping to raise the money to fund a renewed educational system designed to address the needs of all students.

DISCUSSION

Educational testing has a long, and in many ways, proud history in the United States. Psychometric theory may be the most enduring of contributions to education to have emerged from the field of psychology. It has provided education with a powerful set of tools, which if used correctly, can lead to improved educational policies and practices. This psychometric tradition has in recent years gathered a growing crowd of detractors who are opposed to the underlying concept of the normal distribution of traits or the so called "bell curve". The argument that has developed is that the normal probability curve that underlies traditional psychometric theory is obsolete, having been replaced by the philosophic tenet that "all children can learn." In this greatest of democratic societies, in an age of increasing diversity, it is at the least politically incorrect, if not an act of educational heresy to challenge or question this precept. It is not the precept itself that deserves challenging, but the extrapolation of the precept to "any child is capable of learning any subject matter at a high level of mastery." This untested assumption, however egalitarian and desirable it may be, has not yet crossed over from the realm of wishful thinking to that of

scientific fact. Surely there are many factors that must come into play for this to be even remotely true. Factors such as adequate home support in the form of nutrition and parental care, parental participation in the child's education, equal access to quality instruction, adequate resources to support the instructional processes, an educational system that does not stigmatize those who take longer to learn a particular type of subject matter, and so on.

The psychometric tradition has made many major contributions to American education and society. This tradition is rooted in the natural sciences and has developed to its present state through generations of research. We know its strengths and its potential and we understand well its limitations. What we are witnessing now in the swing toward performance based assessments is a movement that has no roots in scientific inquiry. It lacks the theoretical underpinnings of psychometry, and instead rests on a somewhat flimsy, if not well intentioned philosophy of intellectual egalitarianism, the validity of which has not been demonstrated. We must refrain from substituting hope for scientific knowledge in our quest for educational perfection or we must face the risk of creating a backlash that will destroy those hopes.

The current PBA movement has grown out of a disdain for NRT theory and practice. Much of the discontentment with NRTs however, stems from misuse, misinformation or misapplication of NRT data. A simple example is the failure to recognize that NRT survey achievement tests are not designed to be used as individual diagnostic tests. To many educators who have had little or no psychometric training, a test score is a test score. There is little concern for such technicalities as the standard error of estimate and the implications it may have for using an individual test score in isolation from other data. Misuse of NRT data, founded on ignorance, has contributed to much of the negativism toward NRTs. This negativism has largely emanated from the ranks of those instructional experts whose own training in and understanding of norm-referenced testing theory is quite limited.

The PBA movement may contain the seeds of its own destruction. Without a scientific core it cannot, will not and should not survive. Therefore if the children of today and tomorrow, those who will be called upon to lead the world community through its most difficult challenges are to appreciate the benefits of its promise, then it is crucial that the scientific research needed to better understand how to create sound, valid and useful performance assessments beginning now on a broad scale. Such research is essential to answer questions of the appropriateness of such concepts as validity and reliability to these assessments. Research is needed to validate our assumptions about the most appropriate uses and

interpretations of the data generated by these assessments. Research is also needed to enable and support the development of appropriate and rigorous technical standards for these assessments.

Scientific research by itself however, will not automatically result in improved assessment practices. For such improvement to occur we need a mechanism for moving the research results into the practical arenas of the classroom, the superintendent's office, the board room and increasingly the venues of the proliferating citizens' advisory committees. A federally sponsored resource bank of scientifically proven assessment practices, based on widely accepted technical standards, could provide such a mechanism.

The rush to reform education through its assessment techniques must not renounce ninety years of scientific knowledge. Instead its leaders must find ways for modern psychometric theory to work collaboratively with the curriculum and instruction methodologists to develop a rigorous set of standards to guide the development of these new assessment techniques. Teachers, testing experts and curriculum designers must work together collaboratively to explore this new territory and chart carefully its potentially perilous terrain. If they don't, educational assessment and the promise it holds for significant reform may founder and breed not new conquests and greater heights of achievement for all, but breed instead disappointment, failure and a retreat to the past. Currently, teachers and curriculum leaders in our public schools lack the expertise, and I believe in most instances, the interest in establishing the technical rigor that will be critical to the success of the PBA movement. The scientific expertise and energy for this work must therefore be supplied by the research and testing professionals. These are the people who must formulate the research paradigms and lead the way in conducting the research that will ultimately answer the many questions about PBAs. This research has already begun as is evidenced by the recent works of Richard Shavelson, Bob Linn, Eva Baker and the entire staff of the Center for Research on Educational Standards and Student Testing (CRESST).

A recent article in Education Week (Rothman, 1993) revealed an effort sponsored by the Ford Foundation, to establish principles to ensure educational equity in the newly emerging assessments. The guidelines of this initiative include:

- > field testing new assessments with a diverse population
- > standards and tests that reflect the skills and knowledge for which they are intended (validity)
- > variety in the options available for students to demonstrate their knowledge and skills
- > policies that list the programs to be replaced by the new standards and assessments

The Ford Foundation group called for an expert panel to have oversight responsibility for the development of new assessments, much like the Underwriters' Laboratory or Consumers' Union. I applaud this group's efforts and endorse the concept of such an oversight panel.

Education is often criticized for, among other things, being susceptible to faddism in curriculum, instruction and organizational theory. Whether the movement toward performance based assessment is to be another short-lived fad or not will depend on a myriad of factors. Among these are: how the politicians respond to this phenomenon with policy and funding, how well traditional psychometric theory can demonstrate its usefulness, how effectively a symbiosis can be developed between psychometric research and the curriculum/instruction community, how teacher training is handled, how well PBA advocates can meet the technical and cost effectiveness challenges of PBAs, and how well the PBA approach is presented to and subsequently accepted by a public long steeped in the tradition of comparative status measures. How these issues are resolved is a potentially serious matter, for they could result in yet another failed fad or significant improvements in educational achievement for our country's children.

Worthen (1993) has identified 12 major issues that must be resolved if the movement toward alternative assessments is to succeed in living up to the hopes and desires of its most fervent advocates. These issues are:

- > conceptual clarity
- > a mechanism for self criticism
- > support from well informed educators
- > technical quality and thoughtfulness
- > standardization of assessment judgements
- > ability to assess complex thinking skills
- > acceptability to stakeholders
- > appropriateness for high stakes assessments
- > feasibility
- > continuity and integration across educational systems
- > use of technology
- > avoidance of monopolies

I add to this list the following:

- > potential for ethno-cultural bias
- > aggregation and reporting
- > applicability for making policy decisions

While it was beyond the intended scope of this paper to examine these issues individually, many of them have been recognized in my alternative scenarios of the future of assessment. It is clear that any

innovation fraught with so many important issues and faced with the burden of living up to such great promises, must be subjected to the scientific scrutiny of thoughtful research prior to being implemented on a broad basis. It is essential to our understanding of how to make the best use of these assessments that we conduct such research on a broad and intense level, for without the knowledge gained from such research we are, in all probability, consigning performance assessment to the dustbin of history and thereby losing out on a great opportunity to improve the quality of American education. We must learn to go slow to go fast.

References

Rothman, R. (1993) Week, March, -22, 1993. Group Outlines Principles for Equity in New Assessments. Education

Worthen, B. R. (1993) Critical Issues That Will Determine the Future of Alternative Assessment. Phi Delta Kappan, JA (6), 444-454.

Symposium II

Objectifying the Subjective: Rubrics, Scoring Guides, and Other Ways of Knowing

M. Kevin Matter, Organizer
Cherry Creek (CO) Schools

Performance assessments are an increasingly popular form of measurement in public schools. This symposium provides some current information on the state of the art in public schools, important considerations in the development of performance tasks and rubrics, and what education can learn from military applications of this type of assessment. Carole Perlman reports on a 1992 survey of NAID members on the use of performance assessments in their districts. Lauress Wise describes how rubrics developed for job performance in the military may be applied in the schools. Judith Arter presents some key components to successful development of Scoring rubrics, while Linda Carstens provides a sourcebook of ideas and samples of rubrics from a variety of teachers. Joe McDonald provides insightful reactions on these papers and the difficulties inherent in performance assessments.

Quantifying Quality: Results of the NATD Performance Assessment Survey

Carole Perlman
Chicago Public Schools

In fall 1992 the National Association of Test Directors (NATD) surveyed its members on their involvement in the development of performance assessments and scoring rubrics. Follow-up questionnaires were mailed and responses were received from 64 members, about a third of the total membership. The purpose of this paper is to examine the extent to which members have developed performance assessments, how they went about doing so, the advice they would offer others who are developing performance assessments, and the nature of the scoring rubrics they developed.

Who responded? How many have developed performance assessments?

About three-quarters of the respondents (49 of the 64, or 76.6%) are employed by local educational agencies (LEAs). The remainder are divided among colleges and universities, educational service districts, state educational agencies (SEAs), consultants, and other types of organizations. Because of their diversity and small sample size of each of the non-LEA subgroups, most of the analyses will focus on either the entire group or on the LEA respondents. Of the LEA group, 21 of the 49 respondents represent school systems with an enrollment of more than 35,000. The enrollments of the districts represented range from 2,800 to 618,000.

Table I gives a breakdowns by organizational affiliation of the respondents who have and have not developed performance assessments. Slightly fewer than half the respondents (43.8%) report that they have developed performance assessments. It's not clear how well that percentage generalizes to other NAM members, but the actual percentage may well be lower, if one assumes that people who have developed performance assessments are more likely than others to fill out a questionnaire with the heading "NATD Performance Assessment Survey."

In what subjects and for what grades have performance assessments been developed?

Writing is by far the area in which the greatest number of performance assessments are being developed (see Table 2). In fact, of the 28 members who had developed performance assessments, 24 had developed writing assessments. Reading and mathematics run a distant second and third. There was curiously little development reported in some of the areas that have been widely regarded as lending

themselves well to performance assessment: science, social studies, the fine arts, listening, speaking and foreign languages. This doesn't necessarily mean that performance assessments are not conducted in those areas, but it might indicate that development of such assessments is taking place. at the school or classroom level (NAID members employed by LEAs are generally part of their district's central administration) or that schools are using assessments from other sources (e.g., state assessments, assessments purchased from publishers).

Figure 1 gives frequencies and the mean number of subjects in which assessments were developed by LEAs. Of the 22 LEA respondents who developed performance assessments, 9 developed assessments in only one subject area and 8 developed assessments in two subjects. Just under one-fourth report developing assessments in three or more subjects. The average number of subjects was 0.94; considering only the respondents who had developed at least one performance assessment, the average number of subjects was 2.09.

The grade levels for which performance assessments are developed vary by subject area (see Figure 2). For the LEA respondents, most of the development in reading and all of the math is concentrated in grades K-3. Writing and speaking assessments are more evenly distributed across the grades.

Does school system size affect the number of performance assessments developed?

Table 3 shows the number and percentage of large and smaller school systems that have developed performance assessments. Performance assessments in at least one subject were developed by 42.9% of the larger systems (N = 21) and 46.4% of the smaller ones (N = 28). Although about the same proportion of large and smaller school systems developed math performance assessments, larger districts were more likely than smaller ones to have developed writing assessments and less likely to have developed reading assessments, though the differences were not statistically significant.

Figure 1, which gives information on the number of subjects in which the respondents developed performance assessments, presents data for large and smaller school districts. The mean numbers of subject areas are almost identical for the two groups.

For LEAs, are performance assessments developed primarily at the classroom, school, or district level?

Twenty-one members affiliated with LEAs responded to this item. The classroom and district levels were cited with equal frequency (57.1 %) as the primary locus for development of performance assessments.

Development at the school level was mentioned only half as often (28.6%). The only difference between large and small districts was that respondents from the smaller LEAs were more likely to say that the classroom was a primary locus of development. It is not clear whether less development is going on in the classrooms of large school systems or whether the respondents are less likely to know about such development in a very large school system.

Which LEA office or department has responsibility for developing performance assessments?

Of the 19 members who responded to this question, five indicated that the responsibility rested with testing, evaluation, and research staff and three cited curriculum and instruction. Eleven members said that responsibility was shared by the two offices.

What is the role of NATD members in the development of performance assessments?

Figure 3 summarizes members' responses to the question, "What is your role with respect to performance assessment (e.g., are you the primary developer, technical consultant, trainer, etc.)?" The most frequently mentioned role is that of technical consultant (73.9% of LEA respondents and 70.0% of all respondents). About a third of the respondents serve as a trainer, coordinator of performance assessment development or data gatherer/analys reporter. Other frequently mentioned roles were primary developer and developer. Here is a sampling of how members see their roles:

All of the above plus collaborator, broker for workshops and training sessions, etc. [Educational Service District]

All of the above. I am in charge of the development and also serve as a technical resource. (We have others, but they are outside consultants.) I also train teachers to use the assessment system. [University-based assessment center that develops early childhood performance assessments]

Supervise development of systemwide performance assessments; scan and report assessment results; serve as technical consultant to central and school staff-, conduct inservice, training on performance assessment and portfolio development for teachers and principals. [LEA, 411,000 students]

Oversee all administration of assessments in district; work with Curriculum Department to develop events/ open-ended questions; be knowledgeable and oversee writing and math portfolios, alternative portfolios for Special Education and primary portfolios for K-3. [LEA, 90,000 students]

Technical consultant; coordinator of district efforts; district representative and curmudgeon to the state testing people. [LEA. 65,000 students]

I beat the drum. Budget constraints severely impact our ability at present to pursue performance assessments; however, there is great enthusiasm in our schools. [LEA, 44076 students]

Supervisor of area...[Another person, who has primary responsibility, is assisted by graduate students.] I guess I am mostly a cheerleader. [LEA, 44,000 students]

Conceptual leader, trainer and consultant. [LEA, 31,000 students]

Technical consultant...I also conduct all scoring training workshops for teachers. We have mentors assigned to me that are becoming "assessment trainers/experts." I help candidates write assessment projects each year. [LEA, 20,000 students]

I did the project with the assistance of teachers ... [whom] I hired to write items. [LEA, 18,000 students]

We are beginning to adapt/develop performance assessments as part of our comprehensive evaluations. [Non-profit organization engaged in program evaluation]

How were assessments and rubrics developed?

Twenty-three members (20 from LEAs, 2 from SEAs, 1 from a college) described the process by which their performance assessments and rubrics were developed. Except for one LEA staffer whose school system purchased performance assessments from a publisher, the development process showed remarkable uniformity. Tasks and rubrics were generally developed by teachers or curriculum staff (or, much less frequently, by measurement specialists with input from those groups). Teachers were an integral part of the development process in nearly all cases. About a third of the LEAs reported that they adapted existing rubrics obtained from their states or other outside agencies. Many members reported a painstaking, iterative process of consensus building, reviews, pilots, analyses, and revisions. Some of their responses are given below:

Writing objectives were identified by English teachers. Prompts were written by committee of English teachers and field tested. Sample of field test papers was used by committee to draft preliminary scoring rubric. Prompts are refined for districtwide administration. A random sample of papers from a stratified random sample of schools is pulled and used to refine scoring rubric

and prepare training packets for readers. [LEA with 618,000 students, writing assessment at grades 7-12]

Literature-Based Writing Process Assessment developed by teachers over the past four years. Revised rubric in 1990 to use a variation of the NWEA 6-trait 5-point rubric for writing. [LEA with 32,000 students, writing assessment at grades 1-8]

We recruited a large group (approximately 40 teachers and administration) to design the assessments. After spending about half the school year in research and assessment training, we decided that the majority of our work initially would focus on criteria writing. Tasks are more easily found, borrowed, or purchased from other sources. Deciding what to judge and what to look for in student work and behavior would be the most important first step. We decided that standards could only be set after the tasks and assessment are field-tested and validated, at which point the question of "how good is good enough?" can be addressed. [LEA with 32,000 students, writing and critical thinking assessments at grades 9-12]

One individual with content expertise for preliminary draft-reviewed and revised by all appropriate grade level instructors-final draft small committee. [LEA with 12,500 students, reading assessments at grades 1-6, writing assessments at grades 1-6 and 12]

Teachers developed them based on Texas essential elements and district objectives. They were piloted and revisions made. Instruction and assessment were combined. [LEA with I&OW students, listening and speaking assessments at grades 1-6]

Committee of teachers at grades K-3: (1) researched into performance assessment and best practices currently being used; (2) established outcomes for primary mathematics; (3) selected tasks that validate these outcomes; (4) designed rubrics for scoring; (5) field tested and revised based on teacher comments. [LEA with 8,413 students, math assessments at grades 1-2]

How did NATD members investigate the technical quality of performance assessments?

Eighteen of those surveyed (62.1 % of those developing performance assessments) indicated that they had investigated the reliability and/or validity of their assessments. Of the 14 respondents who gave specific descriptions of the studies they conducted, ten had measured interrater reliability. Validity studies were mentioned much less frequently. The ways of investigating validity included:

- > comparison of scores with final course grades and a study of whether students enrolled in higher level courses scored higher [college]
- > correlation with objective writing assessment [LEA]
- > disaggregation of results by gender, ethnicity and language classification [LEA]
- > multi-faceted analysis of ratings across task, rater and content using item response theory [SEA]
- > gathering evidence recommended by Linn, Baker and Dunbar (1991) regarding consequences, generalizability, fairness, cognitive complexity, meaningfulness, content quality and coverage, and cost justification. [university-based assessment center constructing prekindergarten-grade I assessments]

Is student performance tied to any significant consequences?

Except at the high school level, members report that performance assessments are generally low stakes tests, at least from the student's point of view. Of the 18 LEA respondents, three said that students must currently pass a performance assessment in order to graduate, two reported that a similar requirement will be implemented soon and four indicated that passing a performance assessment was needed for certification of competency at a high school grade other than 12. Four members reported that performance assessments were used to determine placement into courses or special programs (e.g., remedial, gifted), three indicated that results will be linked to school accreditation, and two said that performance assessments were used in making promotion decisions.

What types of rubrics are used?

Nineteen members submitted one or more scoring rubrics, not all of which were developed by the member or his/her organization. Not surprisingly, most of the rubrics are for writing assessments. Of the writing rubrics, 15 were analytical (i.e., separate scores are assigned for specific features of the writing), four were holistic (i.e., there was a score for overall performance) and six used both analytical and holistic ratings. Only one member submitted writing rubrics that were prompt specific. Most of the other rubrics submitted were for assessments administered in the early elementary grades. These include tasks in a variety of subject areas, progress reports, journals and student self-assessments. The rubrics for writing, reading, mathematics, listening, speaking and science are described in the appendix. Copies of some tasks and rubrics submitted by NAID members are shown in a companion document.

Which procedures have proven successful for developing assessment tasks and rubrics?

When asked what advice they would give to others who need to develop performance assessments, nearly all of the 20 respondents (all but two from LEAs) mentioned getting extensive teacher input and allowing enough time. Here are some of their comments:

Consider test development to be formative and subject to much revision. For the rubric: it is important to reach a consensus (but recognize that there will always be outliers!). [college]

Involve a broad base of teachers in the development of tasks and rubrics after "umbrella" district objectives and "standards" have been established. [LEA, 618,000 students]

Teacher input is critical. Allow enough time to pilot and revise as much as necessary. Keep rubrics simple-teachers seem to prefer fewer, more global ratings to a larger number of more detailed ones. A single performance assessment isn't going to provide all the information needed; use multiple measures and consider combining performance assessments with more traditional measures. [LEA, 411,000 students]

Provide initial training for raters until they score reliably each time/session they work. (Even the best get "rusty.") [LEA. 55 PM students]

Don't re-invent the wheel; use a developed model and modify as necessary. Read the literature. Do involve the teachers as local readers; it definitely improves instruction. [LEA, 43,000 students]

Staff development activities and incentives/grants for developing tasks/rubrics for classroom or school use. ILEA, 32 ~OW students]

Need some form of staff development when involving teachers; requires administrative leadership. ILEA, 31,524 students]

Good training and discussion about the relevant dimensions and the range of scores is absolutely necessary to good performance assessment. Tasks seem to be relatively easier to create or adapt once the criteria we established (although the field-testing may show more problems with task selection than we anticipate!). A second hurdle to overcome is the conviction that performance assessments can provide a sufficient amount of information by themselves to document student learning, without evidence from additional, more traditional measures. Time and again teachers want to make decisions about students based upon one writing prompt, for example, in spite of evidence that the information is very narrow. One thing that helped this was

gathering empirical evidence and showing teachers that the amount of error was very large when minimal amounts of data are used. [LEA, 24,000 students]

Top down doesn't work. Collaborative models are best. Teachers are motivated to find new assessments to support areas on our new report card. They have to feel confident when explaining scores/grades to parents. [LEA, 20,000 students]

Each teacher needs ownership and buy-in. Get input from everyone at some stage in the process, even if it's just a final "read and comment" request. [LEA, 18,000 students]

[1] Find balance between assessing via individual interviews and integrating with instruction. [2] Keep it teacher based; trust teachers. [3] Defining criteria is hardest and most rewarding task for teachers who participate. [4] Build consensus. [LEA, 14,000 students]

Teachers, time and district commitment of considerable financial resources needed. [LEA, 13,400 students]

Tests need lots of pilot time for review by people who actually administer them. [LEA, 12 500 students]

We have worked with groups of teachers who then sought input from their peers before, finalizing. [LEA, 5,500 students]

Successful: Lots of involvement from teachers; repeated pilot studies and rounds of revisions (one or two isn't enough!). Unsuccessful: Trying to go too fast! This type of development takes at least twice as much time (and probably more) than development of traditional tests. Teacher training must be comprehensive and ongoing. [university-based assessment center constructing prekindergarten-grade I assessments]

Tasks must be clearly stated and must cover a wide range of abilities and skills; rubrics must be open enough to capture richness of student responses; scoring must represent clear rating scales. [SEA]

Discussion

It is clear that, while there is great interest in performance assessment, development activities (at least at a district-wide level) are not being carried out by most of the respondents and are hardly being carried out at all in subjects other than writing. The reasons for this are unclear-, the cause may have to do with a lack of time or money; insufficient dissatisfaction among decision-makers with existing assessments; the

feeling that performance assessments should not be imposed in schools in a "top-down" fashion; or the possibility that such development is being carried out at either the classroom or the state level. The survey only addressed the issue of whether performance assessments were being developed, not whether they were being used. In retrospect, it would have been better to ask whether performance assessment was being used, and, if not, then why not. It would also be interesting to know to what extent portfolios are being used (only a few respondents mentioned them) and the processes by which portfolios become valid and reliable measures.

"From the Bottom Up" A Sourcebook of Scoring Rubrics Designed by Teachers

Linda Carstens
San Diego City Schools

Use of Performance Assessments

Performance assessment, also called "alternative" and "authentic" assessment, measures actual student performance in a given subject area. A measure of a student's ability to type provides a good example of the difference between traditional standardized testing and performance assessment. A standardized, multiple choice test might ask the student to label the keys of the type writer with the correct symbols or to identify the fingers touching the keyboard when typing a capital 'T'. A performance test would simply ask the student to demonstrate the ability to type by typing something. The aim of performance assessment, then, is to engage students in assessments, that better represent the "tests likely to face them as professionals, citizens, or consumers (Wiggins, 1989).

Scoring within Performance Assessment

Even though it is apparent that alternative forms of assessment must be developed to evaluate student learning, practitioners and researchers will be quick to point out a number of concerns related to scoring.

- > How will the assessments be scored?
- > Can we rely on the professional judgment of teachers?
- > How can we ensure equity within scoring systems?
- > How do we promote quality assessments through an objective, meaningful scoring system?

These questions and others must be addressed before parents, community members, and other educators will have a sense of confidence about performance assessment in general and the reliability of scoring in particular. For in-depth information about scoring, the reader of this Sourcebook is referred to *A Practical Guide to Alternative Assessment* (Herman, Aschbacher, and Winters, 1992) and *Testing for Learning* (Mitchell, 1992).

The Scoring Rubric Sourcebook

This sourcebook is a descriptive collection of scoring techniques and activities being used by various national, state, and local educational organizations. The open format of the sourcebook allows

information to be added as appropriate and available. For easy reference, the type of assessment being scored (such as portfolio or exhibition), the level of the assessment (such as classroom or district), the curriculum area(s), and the grade level(s) are listed at the top of each page. General information about the scoring technique is provided on the body of the page, and the source of information about the assessment and scoring is listed at the bottom of the page. If further information is desired, the reader is encouraged to contact the sources listed or the author (see page iii for address).

PRINCIPLES OF GOOD PRACTICE IN ASSESSMENT

In her article, "An ABC of Assessment", Myra Barrs lists seven principles of assessment that reflect good practice, are informative (for children, parents, other teachers, and wider audiences), and do not distort the teaching and learning process.

1. The assessment of normal behavior in favorable contexts.

Sampling of normal behavior in favorable day-to-day contexts gives a better and more reliable picture of a student's capabilities than does a one-time assessment in an unfamiliar situation. Assessment in familiar situations can address issues such as the value of extended units of work.

2. The importance of assessment across a range of contexts.

This principle stresses the role of context in determining performance, reminding us that performance may vary in different kinds of contexts and that different kinds of tasks call for different skills.

3. The assessment of process as well as product.

Barrs believes that "when teachers switch their focus from the end product of a piece of learning to the actual process of learning, they form a fuller picture of the child as a learner, and are in a better position to intervene intelligently and to support children's learning helpfully."

4. The holistic assessment of complex processes.

Attempts to measure achievement analytically, especially in reading and language, are too crude to be useful and are based on inadequate models of learning.

5. The sharing of criteria with students.

When students understand the criteria being used to evaluate their work, they participate in the process of assessment. As a result, student self-assessment becomes more informed and reliable.

6. The inclusion of pupil self-assessment.

Students should be involved in their own assessment through such activities as keeping records of their work in list or diary form, keeping reflective journals, conferencing with teachers, and providing information for their own summative assessments.

7. Equity in assessment.

"Unless students have equal access to curriculum, they cannot have equal treatment under any assessment scheme," says Barrs.

Source:

"An ABC of Assessment" (1990)

Myra Barrs, CLPE

ILEA/Centre for Language in Primary: Education

United Kingdom

A Sourcebook of Scoring Rubrics

Assessment: Portfolio/Performance

Assessment Level: Classroom,

District, National

Curriculum Areas: All

Grade Levels: K-12

GETTING STARTED: A TWO-POINT RUBRIC

Background The overarching issue for scoring is always how well the response accomplishes the prompted purpose. On demand performance responses are drafts, not final edited works. Even the best responses will typically need some editing to correct minor flaws. Within the framework of accomplishing the task's purpose, the rubric asks the scorer to appraise the response according to how well it exemplifies the ideas, knowledge, and thinking power described in the curriculum frameworks.

The Two-Point Rubric from New Standards Project

The process for arriving at a consensus calibration for a group of scorers using a four or six point rubric begins with sorting responses into two categories: Ready for Revision and Instruction Needed. The basis for this sort is the scorers' professional background And the following two point guide:

Ready for Revision (4,5,6): A response that essentially accomplishes the task or could be revised to accomplish the task. The response contains convincing evidence that the student has learned enough to tackle the task effectively. Given feedback specific to what is in the response, more time and effort, the student would produce a good quality response without additional instruction. The response may contain errors and omissions, as long as the response as a whole has the thinking and substance of a successful response.

Instruction Needed (1,2,3): Not enough accomplished to revise, needs instruction. The response lacks convincing evidence of enough learning to effectively accomplish the task, even with more time and feedback specific to what is in the response. Sustained interaction or instruction that goes beyond feedback on what is already in the response is needed.

Source:

New Standards Project
Learning Research & Dev't Ctr.
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

A Sourcebook of Scoring Rubrics
Assessment, Type: Performance Task
Level of Assessment: National
Curriculum Area: General
Grade Level: Grade 4

EVALUATING STUDENT WRITTEN WORK: HOLISTIC SCORING

Background

Holistic scoring is used primarily to assess student writing or other open-ended student products that do not have one "correct" answer. According to "Assessment Alternatives in Mathematics", holistic scoring is a sorting activity in which teachers first sort student work into three stacks: 1) student work that missed the point all together, 2) student work that was acceptable, and 3) student work that has some special quality. Each of these three groups are then broken down even further into two levels, as shown in the diagram below. Each of the resulting groups can be assigned a numerical score, if desired.

When this method is used, students' thinking, Problem-solving abilities, and communication skills can be assessed without the need for "one right answer".

Additional Information:

"Assessment Alternatives in Mathematics" (1989)

EQUALS I

Lawrence Hall of Science

University of California

Berkeley, CA 94720

A Sourcebook of Scoring Rubrics

Assessment Type: Scoring Techniques

Assessment Level: Classroom, State, National

Curriculum Areas: All

Grade Levels: K-12

EVALUATING STUDENT WRITTEN WORK: RUBRIC SCORING

Background

Like holistic scoring, rubric scoring is used to assess student writing or responses to open-ended problem solving activities. While holistic scoring sorts student products according to general levels of response, rubric scoring rates student work on a specific problem or activity as it compares to a certain pre-determined standard. A rubric, according to "Assessment Alternatives in Mathematics", is a "description of the requirements for varying degrees of success in responding to an open-ended question." Rubrics may be written after student work has been collected or before students begin working on a problem or activity. In classroom settings, it is advisable to share rubrics with students before they begin to work on responses to problem-solving activities or research projects. In this way, students learn to structure their thinking and written work in meaningful ways.

Rubric Development

The California Learning Assessment System (CLAS) developed and used individual rubrics to score each of the open-ended math questions on the statewide pilot last year. These scoring rubrics were written by teachers who discussed appropriate responses for each question. Categories of responses developed by the teachers were similar to those used in the description of holistic scoring mentioned above. CLAS provides the following suggestions for developing rubrics for open-ended math problems.

1. Have your students work the problem.
2. Have your faculty colleagues in mathematics do the same problem.
3. Discuss the problem as a group and sort the student papers into six groups with six being the highest and one the lowest.
4. Discuss the characteristics of the responses and articulate a rating for an exemplary rating (six) response. Articulate rubrics for the other categories.
5. Take a second look at the student work and regroup them as needed.

Source:

"Assessment Alternatives in Mathematics" (1989)

EQUALS

Lawrence Hall, UC

Berkeley Berkeley, CA 94720

A Sourcebook of Scoring Rubrics

Assessment Type: Scoring Techniques

Assessment Level: Classroom State, National

Curriculum Areas: All

Grade Levels: K-12

THE CALIFORNIA LEARNING RECORD

Background:

The California Learning Record (CLR) was adapted from Britain's Primary Language Record (PLR). Both the CLR and the PLR extend their scope beyond traditional student assessment in that: 1) both parents and students are involved in developing and maintaining the record, 2) records can be used to assess both English-only and bilingual students, and 3) records provide teachers with a framework for teaching language and literacy. In California, the CLR is being examined as an alternative to Chapter Year-End Observations (completed at year's end) This section: - gives parents and child a chance to reflect on progress - records the final assessment in all areas of language arts 9 gives teachers a chance to pass along information and suggestions to child's next teacher.

Parts of the California Learning Record:

Part A:

Background Information

(completed first quarter)

Information about the student

including:

- All staff involved with child
- Languages read, spoken, written, understood
- Discussion between parents and teacher
- Record of language/literacy conference with child

Part B:

Child as Language User

(completed during year)

Includes daily records and

observations of child's literacy

development in:

- Talking and listening
- Reading
- Writing

Part C:

Year-End Observations

(completed at year's end)

This section:

- gives parents and child a chance to reflect on progress
- records the final assessment in all areas of language arts
- gives teachers a chance to pass along information and suggestions to child's next teacher.

Additional Parts of the CLR

Also included in the CLR are informal observation and sample sheets that allow teachers to describe significant points in a child's language and literacy development by examining samples of the child's reading, writing, speaking, and listening.

Source:

California Learning Record
Dr. Mary Barr, Co-Director
University of California, San Diego
9506 Gilman Drive a Jolla, CA 92093

A Sourcebook of Scoring Rubrics
Assessment: Portfolio/Performance
Assessment Level: Classroom, District, Nation
Curriculum Area: Rdg/Language Arts
Grade Levels: K-8

National Association of Test Directors

CENTRAL PARK EAST: END-OF-COURSE EXHIBITION

Background Central Park East Secondary School, a member of the Coalition of Essential Schools, is located in New York City's East Harlem. Each student studying Humanities must complete an exhibition to culminate the school year. Excerpts from the student exhibition assignment sheet follow.

This week, you will begin to work on your final exhibition. Your final presentation will be oral, written, and visual. We have studied four time periods: 1) the 1988 elections, 2) the American Revolution, 3) the Civil War, and 4) the Sixties. This is your chance to show us and yourself what you have learned this year. You must choose one activity form each Part. (Note: Only one example is shown from each section of the assignment sheet.)

Part I: Create a dialogue between two famous people from different time periods (for example, George Washington and Malcolm X). The dialogue must include a discussion relating to one or more of the essential questions. Whichever personalities you choose, you must represent their point of view and the historical period in which they lived.

PartII: Literature Choose a character in a particular scene from either "Of Mice and Men" or "The Chocolate War". Decide whether the person is acting from a position of power or powerlessness. In order to do this, you must show the character in a particular situation or scene. Use quotations and examples.

Part III: Visual Draw a poster or make a collage that shows the important ideas of the historical times we have studied.

Part IV: Reflection How did the study of the essential questions make you think about your ability to create change in your own life? Think about a current situation in which you were able or would be able to change something.

For more information, contact:
Central Park East Secondary School
1573 Madison Avenue
New York, NY 10029

A Sourcebook of Scoring Rubrics
Type of Assessment: Exhibition
Level of Assessment: School
Curriculum Area: Humanities

Grade Levels: Middle/Senior High

CONNECTICUT PERFORMANCE ASSESSMENT PROJECT

Background

The Connecticut Common Core of Learning Assessment Project, sponsored by the National Science Foundation, is in the process of developing a statewide performance assessment in science and mathematics. In 1991-92, the system was piloted in approximately 30 schools, and is now implemented statewide. This project has now been expanded to a consortium of six states and the Coalition for Essential Schools.

A Sample Mathematics Task

This grade 10 geometry performance task, "Building a Dog Pen", is intended to extend and assess student understanding of the relationship between the area and perimeter of polygons. The task: Given 80 feet of fence, what is the largest area that can be enclosed to form a free standing dog pen? Given 36 square feet of area, which shape--a triangle, rectangle, square, or circle--uses the most of the 80 feet of fence available for a free standing dog pen? Students work for several days, both individually and in groups, to solve the problem. They then participate in a variety of extension activities and related tasks to demonstrate understanding. Some of these tasks are completed by individual students and some by groups of students.

Scoring

Scoring of the performance task has several dimensions. Whole group criteria look at 1) the performance of the group as a whole in the areas of collaborative process, joint presentations and arriving at or developing content, and 2) performance of the individual student as a member of a group, working in collaboration, and participating in developing content. Individual student criteria include collaboration skills, attitudes, general skills (such as problem solving and communication), and subject specific skills and understandings. Students are also evaluated on the performance task related to the original problem.

Source:

"Samples of CoMPACT Performance Tasks" A Paper Presented at the Annual Meeting of the American Educational Research Association
April 16, 1990 Boston, MA

A Sourcebook of Scoring Rubrics
Type of Assessment: Performance
Level of Assessment: State

Curriculum Areas: Science and Mathematics Grade level: High School

MEASURES OF SUCCESS: AN OPEN LETTER TO PARENTS

Dear Parents,

Assuring success for all students demands that we work together as coeducators to improve our schools. To make meaningful change, we must reexamine the traditional methods we use to teach our children and assess their progress.

Traditional assessments, such as standardized tests, mostly show what children cannot do. Yet these tests are often the only method schools use to report student progress to the community. However, within the classroom, there are other ways to measure student learning. Teacher observations, samples of student work, and student projects can be used to evaluate student success. These "alternative assessments" are an important part of the learning process and focus on what children can do.

Traditional standardized tests do not provide us with information we need to help children succeed at school. In contrast, alternative assessments document progress over time, collect rich information about learning, and provide feedback about the effectiveness of instructional programs and teaching strategies used in the classroom.

As parents, you should be aware that alternative assessments can provide more information about the progress of your children than standardized tests alone. By demanding such methods of evaluation and participating in the assessment process, you can become advocates for education and help to enrich teachers' understanding of the unique needs and strengths of your children. Together, we must challenge the school system to assess skills and knowledge we value in order to promote school experiences that will result in increased success for all our children.

Note: This letter is also available in Spanish.

Source:

Performance Assessment

Development Unit

San Diego City Schools

4100 Normal Street, Room 3133

San Diego, CA 92103

A Sourcebook of Scoring Rubrics

Type of Assessment: General

Assessment Level: Classroom

Curriculum Area: All areas

Grade Level: Elementary
1993 Symposia

MULTI-DIMENSIONAL RUBRIC TO IMPROVE INSTRUCTION

Background

Rubrics which result in the assignment of a single score provide students with little information about the quality of the various dimensions of their work. The writing portfolio project at Stevens Creek Elementary School (Cupertino, CA) provides an example of a multi-dimensional rubric through which students receive feedback about the strength of their work and how it can be improved. The rubric includes separate descriptors of performance related to students developing understanding of specific narrative elements including character, setting, plot, and theme. Although the rubric is still being revised and has yet to be tested using many student work samples, its multi-dimensional design appears to be an extremely useful tool for helping both teachers and students plan for improvement in narrative writing.

An Excerpt from Stevens Creek Elementary's Muld-Dimensional Rubric

arbitrary ←————▶planned one attempt ←————▶revision

- I. Writer works "in the moment" (one attempt); planning centered on drawing and talk.
- II. Child may take advantage of teacher-led group planning but will not stray far from the forms offered; revision focused on convention rather than communication.
- III. Writer makes more consistent use of group planning, and begins to expand on forms offered; revision centers on convention with few changes in content, structure, or voice.
- IV. Planning expanded to include organizational structures, but these may not be referred to once the writing begins; during revision writer makes some communication changes in the beginning or end of the text, writer begins to make revision at word level to refine and enhance the text.
- V. Writer uses prewriting ideas to help guide the piece; planning expands to the use of research; revision includes adding or deleting information to the middle of the text to improve the text; word level revision continues but connects to overarching choices in metaphor.
- VI. Writees early planning and extensive exploration of possibilities result in higher quality first drafts; revision centers on communication as the writer moves toward meaning; may make major alterations and reorganize as writers come to have higher standards for themselves.

Source:

Stevens Creek Portfolio Project

UCLA/CRESST

Graduate School of Education

405 Hilgard Ave.

Los Angeles, CA 90024

A Sourcebook of Scoring Rubrics

Type of Assessment: Portfolio

Assessment Level: Classroom, School

Curriculum Area: Writing

Grade Levels: K-6

STELLAR PERFORMANCES: WHAT DO THEY LOOK LIKE?

California Learning Assessment System (CLAS) Reading is the process of constructing meaning through transactions with text. The following is excerpted from the California Learning Assessment System (CLAS) rubric designed by teachers, which will be used this spring. It describes an exemplary performance based on a six point scale.

Score Point 6

An exemplary performance is insightful, discerning, and perceptive as the reader constructs and reflects on meaning in a text. Readers at this level:

- are sensitive to nuances and complexities;
- make plausible assumptions about unstated causes or motivations;
- differentiate between literal and figurative meanings;
- demonstrate understanding of the whole and how the parts work together;
- connect understanding of text to own ideas, experience, knowledge; to history as participants in a cultural community; or to other texts or works of art;
- draw on evidence from text to generate, validate, expand, and reflect on ideas;
- entertain challenging ideas and explore multiple possibilities of meaning;
- often revise their understanding of a text as they re-read and as additional information or insight becomes available;
- sometimes articulate a newly developed understanding;
- raise questions, sometimes taking exception, agreeing, disagreeing, appreciating, or objecting to text features;
- may test validity of author's ideas, information, and/or logic by considering the authority of the author and the nature and quality of his/her sources;
- frequently suggest ways of rewriting the text, speculating about the ideology or historical biases, sometimes recognizing, embracing, or resisting the ideological position that a text seems to construct for the reader.

From:
CA Learning Assessment System
CA Department of Education
721 Capitol Mail
Sacramento, CA 94244

A Sourcebook of Scoring Rubrics Type of Assessment:
Open-ended Response
Level of Assessment: State
Curriculum Area: Language Arts
Grade Levels: Grades 4, 8, and 10
|

SCORING PORTFOLIOS: LESSONS FROM PITTSBURGH

Background

Pittsburgh has concentrated on the writing portfolio as a process of "production, perception, selection, and reflection" exercised by each student over his or her range of productive work. The Pittsburgh portfolio contains a personally important piece, satisfying and unsatisfying pieces, an entire biography of a work, a free choice, and a mutually agreed-upon choice by the teachers and students.

Pittsburgh Writing Portfolio Scoring System

Dimension 1: Accomplishment in Writing

Meeting worthwhile challenges

Establishing and maintaining purpose

Using the techniques and choices of the genre

Controlling conventions, vocabulary, and sentence structure

Being aware of the needs of the audience (organization, development, Using language, sound, images, tone, and voice Including humor, metaphor, and playfulness

Dimension 2: Use of Resources, Processes, and Strategies for Writing

Effective use of pre-writing strategies

Use of drafts to discover and shape ideas

Use of conferencing opportunities to refine writing

Effective use of revision

Dimension 3: Engagement, Growth, and Development as a Writer

Evidence of investment in writing tasks

Increased engagement with writing Development of sense of self as a writer

Evolution of personal criteria and standards for writing

Ability to see the strengths and needs in one's writing

Use of writing for various purposes, genres, and writing audiences

Source: Pittsburgh Public Schools
West Liberty Center
Pittsburgh, PA 15226

A Sourcebook of Scoring Rubrics
Type of Assessment: Portfolio
Level of Assessment: District
Curriculum Area: Writing
Grade Levels: Grades 6-12

DAILY JOURNAL WRITING

Students receive whole group and individual writing instruction. They write in their journals with assistance everyday. Once a month the students choose their favorite journal entry. Entries are evaluated using a writing assessment checklist.

Scoring Rubric:

S: Secure, works independently. Writes sentences without any help.

D: Developing, completes the task but may need prompts to complete the task. Needs help with writing or spelling words occasionally.

A: Assisted, needs substantial help to complete the task. Unable or unwilling to write own sentences. Copies dictated sentences to own paper.

0: Could not complete the task, could not or would not dictate a sentence or write anything on the paper.

Source:

Stuart Foundations Project
San Diego City Schools
4100 Normal St., Room 3133
San Diego, CA 92103

A Sourcebook of Scoring Rubrics
Type of Assessment: Journal
Level of Assessment: School
Curriculum Area(s): Language Arts
Grade Level(s): Elementary

VALIDITY CRITERIA FOR PERFORMANCE ASSESSMENT

Background The traditional psychometric notions of validity are under review. While these notions were appropriately applied to a behaviorist, bell curve model of learning, they found wanting in a standards-based approach to education, or what Madaus referred to as the "end of psychometric imperialism". A number of new constructs are being proposed. Linn, Baker, and Dunbar have developed criteria that represent this new validity for performance assessment.

Consequences	What are the actual consequences of the assessment? Are there any unintended or adverse effects?
Fairness	Does the assessment consider fairly the cultural background of the students? Have all students had equal opportunity to learn the targeted skills?
Transfer and Generalizability	Will the results support accurate generalizations about student capability? Are the results reliable across raters, and consistent in meaning across locales?
Cognitive Complexity	Does the assessment require students to use complex thinking and problem solving skills?
Content Quality	Is the assessment content consistent with the best understanding of the field? Does the content reflect important aspects of the discipline?
Content Coverage	Does the set of assessment represent key elements of the given curriculum?
Meaningfulness	Do students find the assessment realistic and worthwhile? Are the tasks similar to what a professional in that field would do?
Cost and Efficiency	Are the results worth the cost and time to obtain them? Is the assessment seen as part of teaching and learning?

Source:

UCLA/CRESST Graduate School of Education
405 Hilgard Avenue
Los Angeles, CA 90024

A Sourcebook of Scoring Rubrics

Assessment Type: General

Level of Assessment: All

Curriculum Area: General

Grade Level: K-12

STUDENT REFLECTION: GUIDELINES AND EVALUATION

Background

In the Sophomore House at San Diego High School, students are required to use four types of assessment to develop their skills of evaluating their learning: written assignments, oral presentations, reflective writing, and exit exhibition.

REFLECTIVE GUIDELINES

Physics: How has studying physics changed the way you view the world? Possible topics you may choose to discuss include:

- views of yourself or your abilities
- understandings about the world
- views about science
- attitude toward other people and their abilities and skills
- opinion about a science-based career for yourself
- how your work in science can relate to working or learning in another subject
- other topics which relate to the class and your experiences

English (Part 1): "It's my life and I can do what I want, can't I? Use quotes and details from the literature we've read to support your opinion.

(Part II): Re-read your opinion about "It's my life and I can do what I want, can't I?" which you wrote at the beginning of the school year. What differences and similarities do you see in your opinion?

Math: Write about a significant experience in your math class this year and tell how that experience has changed the way you view mathematics.

Evaluation Process for Reflections:

Special Quality:	(6) Exemplary
	(5) Competent
Acceptable:	(4) Minor flaws
	(3) Serious flaws
Missed the Point:	(2) Begins only
	(1) Unable to begin

Source:

Performance Assessment Development Unit
San Diego City Schools
4100 Normal Street, Room 3133
San Diego, CA 92103

Type of Assessment: Reflection

Assessment Level: Classroom, School

Curriculum Areas: All

Grade Levels: High School

ALTERNATIVE REPORT CARD

At O'Farrell Community School: Center for Advanced Academic Studies, the academic year is divided into six thematic units. Teachers complete student progress reports at the end of each of these six-week units of study. The progress reports for each unit are written to address the concepts covered in that unit. The news that six different progress reports are created by teachers each year. In addition to providing information on student performance in the core subjects of mathematics, science, social studies, and language arts, the progress report addresses the areas of physical education, discovery, projects related to the unit theme, personal and social responsibility, and participation in community service. The sample progress report pages shown below are for the core subjects in Unit 2-Uniqueness and Commonalities.

~~0-Made No Attempt, 1-Partial Completion, 2-Full Completion, 3-Exemplary Work~~

HUMANITIES

- Presents elements of a culture from an ancient civilization
- Compares and contrasts ancient civilizations and American cultures in essay form, demonstrating correct knowledge of sentence structure and capitalization rules
- Participates in a Socratic seminar
- Demonstrates media center research skills
- Writes a descriptive essay
- Presents an oral demonstration of student's gift/skill
- Participates actively in language arts class
- Participates positively in social studies class

TECHNICS

- Measures and estimates length, mass, temperature, and volume using metric and English measurements
- Demonstrates the ability to solve an open-ended word problem
- Collects, organizes, forms conclusions and reports on data
- Maintains a math notebook
- "Packaging myself" project
- Demonstrates knowledge of safety practices and proper use of selected lab equipment
- Conducts one controlled experiment to find a solution to a problem or question
- Demonstrate an understanding of the periodic table of elements and is able to locate 10 elements and describe basic physical properties of each
- Designs a classification scheme for a selected group of objects
- Maintains a science notebook/file,
- Participates positively in science class
- Participates positively in math class

The Humanities section is signed by the Language Arts and Social Studies teachers. The Technics section is signed by the Mathematics and Science teachers. Both sections have space for teacher comments.

Source: Curriculum and Instruction Committee
O'Farrell Community School:
Center for Advanced Academic Studies
6130 Skyline Drive
San Diego, CA 92114-56"

SCORING RUBRICS FOR PERFORMANCE TESTS: LESSONS LEARNED FROM JOB PERFORMANCE ASSESSMENT IN THE MILITARY

Lauress L. Wise
Defense Manpower Data Center

Over the past several years, there has emerged an increasingly universal consensus that: (1) many of the skills we want our schools to impart to young people are not well measured by traditional multiple-choice tests, and (2) universal reliance on multiple choice tests encourages teaching and study habits that are at odds with deeper educational goals. As a consequence, educational assessments at the national, state, and local level are incorporating a wide variety of "performance-based" assessment strategies designed more to assess proficiency at practiced skills than to measure simpler factual knowledge.

Performance-based assessment is a somewhat new endeavor in educational research, but it has been studied for some time in other arenas. Measurement of job performance has long been an area of concern for industrial psychologists where strategies such as "work sample" assessments are not new. For the past ten years, industrial and organizational psychologists working for the Department of Defense have been engaged in an unprecedented effort to develop high fidelity measures of job performance for use in validating job selection procedures and standards (Wigdor & Green, 1991). This effort, known as the Job Performance Measurement (JPM) Project, provides a number of lessons that may be useful in the development of performance-based educational assessments. The general goal of this paper is to present information on the approach to developing and scoring performance exercises used in the JPM Project and to suggest lessons that might also be useful in an educational arena.

Background

The JPM Project was a coordinated effort involving each of the Armed Services, overseen by the Office of the Assistant Secretary of Defense for Force Management and Personnel (OASD-FM&P). Each Service selected a sample of jobs to represent the broad domain of occupational specialties available to their recruits. Although the Services organize their entry jobs somewhat differently, four broad categories may be used to categorize jobs in terms of their cognitive requirements: (1) Mechanical, (2) Administrative, (3) Electrical/Electronic, and (4) General. Job classification systems used by the Air Force and Marine Corps use these four categories. The Army further divides the General category into combat, technical, communications, operators, and general maintenance. The Navy also has further subdivisions within the Mechanical and Electrical families. The sample of jobs selected by the Services

represented the different job categories used in each of their selection and classification systems. Table 1 lists the jobs included in the JPM project by Service and by job category.

Hands-On Performance Tests (HOPT) Extensive job analyses were performed for each of the jobs included in the JPM study, resulting in lists of roughly 500 to 1,000 "tasks" identified in training materials, job-specific manuals, or occupational surveys. Tasks in the initial task lists varied considerably in detail and complexity. Simpler tasks were combined and more peripheral tasks were eliminated to narrow the initial list down to a revised list of from 150 to 250 tasks. Subject-matter-experts (SMEs) rated the criticality, difficulty, and frequency of each of the tasks and sorted them into separate content groups. A final sample of tasks was selected from the revised list for each job based on the information provided by the SMEs. The final number of tasks varied somewhat by Service. The Army selected 15 tasks for each job. The Marine Corps project included as many as 35 distinct tasks for one job. The final sample for each job was designed to emphasize the most critical and frequently performed tasks, to cover each of the different content areas identified in the task sorting exercises, and to avoid tasks that were so easy or difficult that little information on individual differences could be gained. For some jobs, an assessment of testing feasibility also was used in deciding among equally relevant tasks.

Once a set of tasks had been selected, scoring rules were developed for each of the target tasks. Critical steps or behaviors were identified in consultation with SMEs. For the most part, each step or behavior was scored as "GO" if it was performed successfully and "NO-GO" if it was not. In some cases, more quantitative information (e.g., typing speed, time for completion, number of targets hit) was also incorporated. In most cases, the total score for the task was the percentage of steps/behaviors performed successfully. Figure I shows an example of a scoring sheet for one of the truck-driver tasks.

In addition to developing the specific scoring rubric for each task, the project team developed a program for training scorers on the use of the scoring rubric. Generally, tasks were organized into four to eight stations, with a different scorer assigned to each station. Scorers received up to two days of training in scoring the tasks for the station(s) to which they were assigned. Special studies were also conducted that included use of multiple scorers at each station and planned rotation of scorers through all stations.

Table 1

Occupational Specialties Included in the Job Performance Measurement Project

<u>Service</u>	<u>Job Family</u>	<u>Occupational Title</u>
Army	General	Infantryman
	General	Cannon Crewman
	General	Tank Crewman
	General	Radio Teletype Operator
	Administrative	Medical Specialist
	Mechanical	Light Wheel Vehicle/Power Generator Mechanic
	General	Motor Transport Operator
	Administrative	Administrative Specialist
	General	Military Police
Navy	Mechanical	Machinists Mate
	General	Radioman
	Electrical	Electronics Technician
	Electrical	Electrician's Mate
	General	Fire Control Specialist
	Mechanical	Gas Turbine Technician, Mechanical
Air Force	Mechanical	Jet Engine Mechanic
	Mechanical	Aerospace Ground Equipment Mechanic
	Administrative	Personnel Specialist
	Administrative	Information Systems Radio Operator
	General	Air Traffic Control Operator
	General	Aircrew Life Support Specialist
	Electrical	Precision Measurement Laboratory Equipment Specialist
		Avionic Communications Specialist
		Electrical
Marine Corps	General	Infantry (Ridleman, Machinegunner, Mortarman, Assaultman, Unit Leader)
	Mechanical	Automotive Mechanic
	Mechanical	Helicopter Mechanic

Figure 1

Example of Hands-On Performance Test Score Sheet

Instructions to Examinee

"During this test you must drive the tractor and semitrailer through the course. (Explain layout of course.) You do not have to perform PMCS [preventative maintenance checks]. You must perform each maneuver without assistance from me or the scorer assistant. You may adjust the seat and mirrors before we begin the test. Tell me when you am ready.

Operate Tractor and Semitrailer Task Score Sheet

<u>Steps</u>	<u>GO</u>	<u>NO-GO</u>
1. Selected first gear position.	_____	_____
2. Maintained RPM (did not allow vehicle to lug, jerk, or stall).	_____	_____
3. Shifted gears without grinding. (Mark N/A for M915)	_____	_____
4. Drove without riding clutch. (Mark N/A for M915)	_____	_____
5. Kept both hands on the wheel (except when shifting).	_____	_____
6. Used hand over hand steering when turning.	_____	_____
7. Passed restricted roadway without striking roadway markers.	_____	_____
8. Followed. serpentine roadway.	_____	_____
9. Passed serpentine without striking-barriers.	_____	_____

NOTE TO SCORER: Do not allow soldier to stop before entering the restricted roadway and serpentine. Soldier should maintain 10- 15 MPH going through both obstacles.

Rating Scales, An independent approach to job analyses was used in developing the performance-rating scales. The critical incident technique (Flanagan, 1954) was used to elicit examples of specific behaviors judged by SMEs to be particularly effective or ineffective. Hundreds of incidents were collected for each job, edited into a common format, and clustered into 8 to 15 categories. SMEs then participated

in "retranslation" exercises where the edited incidents were matched to preliminary category descriptions and rated with regard to the level of effectiveness that they exemplified. Categories were combined or eliminated where there was some ambiguity about the matches of incidents to categories. Incidents for which there was high agreement as to the category and level of performance they exemplified were considered for use as scale anchors. A separate rating scale was constructed for each of the final incident categories. For each scale, one or more incidents were summarized to describe particularly effective performance (scale levels 6 and 7), particularly ineffective performance (scale levels 1 and 2), and average performance (levels 3, 4, and 5). In some cases there were not many incidents at the middle levels of performance, so the behavioral summaries for the middle performance level were synthesized to be consistent with the summaries at either end of the scale. Figure 2 provides an example of one of the rating scales that resulted from the process.

Figure 2.

Example of Behaviorally Anchored Rating Scale

Using Maps/Following Proper Routes

How effective is each soldier in securing proper maps as needed-, becoming familiar with routes ahead of time when appropriate; using maps effectively; following prescribed routes; and arriving at commitments on time?

<p>Sometimes falls to secure or use; may not be able to read maps properly; is often late in reaching designated location.</p> <p>Sometimes fail to plan route ahead of time; may take unplanned route.</p>	<p>Almost always secures maps if needed and uses most maps effectively; usually completes commitments on time</p> <p>Almost always plans route ahead of time; generally becomes familiar with route before commitment; usually follows planned route.</p>	<p>Always obtains proper maps when needed; is able to use grid coordinates to reach even hard-to-find sites; always completes commitments on time.</p> <p>Always plans and becomes familiar with route before commitments; always follows planned route.</p>
<p>1 2</p>	<p>3 4 5</p>	<p>6 7</p>

In addition to developing well-anchored behavioral scales, the project staff also developed a rater-training program (Pulakos & Borman, 1986). This training program was designed to reduce halo effects

and to promote more accurate comparisons among those rated, building on the results of previous research on similar programs (Borman, 1979).

Wise (1992) suggested some lessons for the design of educational assessments that might be drawn from the JPM endeavor (see Appendix A). The remainder of this paper focuses on the issue of how assessments are scored.

Lessons Learned

The lessons learned from job performance measurement about scoring rubrics may be organized under three general questions:

1. How were the assessment exercises designed or selected?
2. What different types of scoring rubrics were used with the exercises that were selected?
3. How was the adequacy of these rubrics evaluated?

“Lessons learned” relevant to each of these three questions are described in the remainder of this paper.

Development of Performance Tasks

The really important question for performance assessments is not so much how to score as what to score. If inappropriate exercises are administered, the best scoring rubrics in the world will not yield a valid assessment of the target ability. The steps used in the JPM project to identify sets of performance tasks provide an example of how meaningful performance tasks might be derived, although some elements of this approach may not translate easily into the educational assessment environment. A brief description of the steps used in developing the hands-on tests is provided here.

Careful specification of the domain to be assessed The development of performance tests for military jobs was relatively simple because results from extensive job analyses were available and because of a close alignment of job training and job performance. A "manual" was available for each job listing all of the tasks that had to be performed. Training material also was available for each-of the tasks in these manuals. Thus, the domain of "job " had already been divided into a discrete set of behaviors or tasks, and it was possible to possible to create an essentially exhaustive list of these elements that constituted job performance.

Involvement of subject-matter-experts, (SMEs). After some editing to even out the level of detail encompassed by different tasks, surveys were conducted to assess the frequency and criticality of each

task and to develop a grouping of tasks into related clusters. Both trainers and supervisors who had extensive knowledge of the domain of interest were available. For each job, there was a school designated as the "proponent" for that job, which provided input to the task definition process and also signed off on the final results.

Systematic sampling from the target domain Sampling was then conducted within each cluster to ensure coverage of all of the different "types" of tasks with priority given to more frequent and more important tasks. In the educational arena, curricular frameworks serve somewhat the same function of specifying the domain to be covered in an assessment, but there is no direct counterpart of the task lists that provided an enumeration of elements in the domain of interest.

Application to educational assessment The curricular frameworks developed for educational assessments serve the general purpose of specifying the domain of performance to be assessed. Continuing efforts are required, however, to achieve better agreement on how to characterize all of the knowledge, processes, and behaviors that constitute the specific elements of these assessment domains. While "SMEs" are plentiful in the educational arena, there is, unfortunately, no central authority that has the definitive word on the content of each curricular area. Professional organizations, such as the National Council of Teachers of Mathematics (NCTM), have made progress in developing curricular frameworks for use with educational assessments, but the process is cumbersome and universal consensus is nearly impossible. Without a reasonably comprehensive specification of the assessment domains, it is very difficult to determine the extent to which scores derived for specific exercises generalize to the larger domain of the assessment.

Development of Scoring Procedures

Two different kinds of scoring procedures were developed in the JPM Project. The scoring procedures for the hands-on performance tests would appear to be most directly relevant to the development of scoring rubrics for educational performance assessments, but the behaviorally-anchored rating scales may actually be more similar to the type of scoring rubrics most commonly used in such assessments. In both cases, several questions were addressed in developing scoring procedures. Each of these questions and their impact on scoring procedures are described briefly here.

Is there a right answer? The first step in developing scoring guidelines for the hands-on tests was to determine the "correct" procedure for performing each task. In the military performance arena, this was a relatively simple endeavor as there were manuals and training materials that provided definitive

descriptions of correct procedures. In the educational arena, life is not nearly as simple. Indeed, the move to authentic performance assessments reflects, in part, a desire to move away from an overly simplistic, binary view of the world where all responses are either correct or incorrect. Consider the following statement from a description of California's science assessment:

"In a performance assessment students are encouraged to demonstrate understanding by conducting an investigation, collecting and analyzing data, and forming a conclusion. These types of assessments have no prescribed answer, but allow for a variety of appropriate student responses, including writing, drawing, and/or manipulation of data. In order to accommodate a wide range of responses, as well as to encourage the evaluating of the students entire thinking process, holistic scoring guides or rubrics were developed for all tasks" (California State Department of Education, 1993).

Another way of considering the issue of one correct, versus many nice, answers is to ask whether we are assessing performance of a procedure that is specifically taught. In the job performance arena, students were taught to follow and encouraged to practice a specific procedure in performing each task. Part of the paradigm shift in educational assessment has been a desire to develop tests that teachers could and should teach to, and to encompass tasks that students might specifically practice. In reality, however, the new educational assessments tend more toward novel tasks that require students to generalize from specific knowledge and procedures that they have been taught. Consider a common beginning to writing assessment prompts: "Compare and contrast Did anyone ever teach you a specific procedure to be followed in responding to such a task? How about procedures for responding to a prompt such as "Describe your favorite music?"

Where there are many correct ways of responding to an assessment exercise, the performance-rating scales be more relevant for educational assessments than the hands-on tests. The critical incident technique used in developing anchors for these scales might be used in developing examples of particularly effective or ineffective responses. Perhaps scoring procedures similar to those used with diving or ice skating competitions might be developed where guidelines indicate how many points are to be subtracted for various kinds of defects in performance. The hands-on scoring procedures might be viewed from this perspective. The "NO-GO" marks for specific steps were a form of "points off" for bad behaviors.

Are we more concerned with process or output? A second question addressed in developing scoring procedures is whether we want to measure adherence to the procedure that is followed or to judge the

quality of the output that is produced. The hands-on performance measures from the JPM Project included both types of criteria. In the example shown in Figure 1, some of the scoring elements, such as "keeping both hands on the wheel" reflected adherence to procedures that had been taught, while other steps, such as "shifted gears without grinding" reflected outcome more than process. In many cases where there is a clearly correct procedure, judging the process and judging the outcome are equivalent. In the educational arena where there are not always correct procedures, we are forced to rely more on judgment of outcome or product. This may be unfortunate, as it makes it more difficult to link assessment results to instruction designed to improve student performance.

Is adherence to prescribed procedures observable? In order for scoring to be reliable, it is important that what is scored be readily observable by the scorer. One of the primary reasons for scoring output rather than process was that output was always observable, while the process used to develop the output may not have been. "Shifting gears without grinding", for example, requires appropriate changes in tension on the clutch as the shift proceeds. The outcome of grinding gears is very easy to observe (hear), while the process used to produce or avoid this grinding is not.

Application to educational assessment In educational assessment, we are often backed into rubrics for "holistic" scoring of examinee "output" because: (1) there may not be agreement on the correct process for producing the output, (2) it is too difficult to make adherence to the process observable, and/or (3) it is simply too costly to make many detailed scoring judgments about each exercise. Nonetheless, there are plenty of examples, such as the "show your work" and "partial credit" approach for mathematics problems, where more detailed scoring of adherence to process can be used. In the JPM arena, the hands-on scoring procedures were generally considered the ideal because they captured most precisely whether the students were following what they were taught and they provided good diagnostic information. It also seems reasonable to propose that a detailed scoring of adherence to process be the ideal for educational assessments as well, particularly where the assessment and instruction are intertwined.

Where it is necessary to fall back on holistic scoring of output, the process used in developing performance rating scales may be useful to consider. This process involves expert "focus" groups given specific prompts to elicit critical aspects of effective and ineffective performance. Similar efforts might be employed to identify differentiating characteristics of good and bad responses to educational assessment exercises.

Evaluation of scoring procedures.

A final question to consider about scoring rubrics is "How do we know a good rubric when we see one?" There has been considerable debate in the educational community about the extent to which traditional psychometric criteria should be used in evaluating performance-based assessments. Some argue that the impact of the assessment on teaching practices and study habits is more important than the information conveyed in specific scores derived from the assessment. One thesis of this paper is that a more detailed scoring of adherence to instructed procedures is generally preferable to a holistic scoring of output. How can we determine whether this is the case?

Much of the debate over psychometric issues reflects an unfortunate concern over the primacy of reliability or validity. Proponents associated with traditional psychometrics argue that assessment results can't be valid if they are not reliable, while proponents of new forms of assessment argue that it does not matter whether it is reliable if it is what we really want to measure (valid). The difference is that reliability is primarily a statistical issue and can be determined analytically, while validity begins with judgment about what is to be measured.

The appropriate criteria for evaluating scoring rubrics will, of course, depend to a large extent on how the resulting scores are to be used. If the goal of the assessment is primarily instructional, the psychometric characteristics of the scores may not be as important as the impact of the assessment on student attitudes, beliefs, and practices. If, however, the scores are to be used diagnostically or in a "high stakes" evaluation of the student, the teacher, or the school system, then both the validity and the reliability of the scores are critical.

When psychometric considerations are important, generalizability theory provides a comprehensive framework for assessing reliability and, to a certain extent, validity (Webb et al., 1989; Shavelson et al., 1990). In contrast to a single reliability coefficient, the generalizability approach tells us the degree to which scores are consistent across scorers and occasions, and across different exercises. In the absence of more ultimate criteria, validity must be judged in terms of the specification of the domain from which the exercises are drawn and the degree to which scores for a sample of exercises will generalize to the whole domain. Specification of the domain of performance to be assessed was where we started this discourse.

Summary

This paper has been an exercise in comparing and contrasting the assessment of job performance as developed by industrial psychologists and new trends in performance-based educational assessments. In many ways, the task of the industrial psychologists was much easier. The domain of behaviors to be assessed was well specified, there was general agreement on how each task should be performed, and there was a close alignment between how examinees were trained and how they were expected to perform on the tests. In addition, one-on-one testing was feasible for the samples used in the JPM study. Nonetheless, several lessons from these efforts are worth consideration in the development of scoring procedures for educational assessments, including the following:

1. Careful specification of the domain of behaviors being assessed is essential to evaluating the adequacy of any particular sample selected for use in an assessment.
2. Scoring elements that assess adherence to processes that are taught rather than just the quality of the output from these processes will have better diagnostic value and, perhaps, greater validity.
3. Scoring procedures must be anchored to observable criteria, so efforts to make adherence to prescribed practices observable are useful.
4. Generalizability theory provides a useful framework for evaluating alternative scoring rubrics.

References

- Borman, W.C. (1979). Format and training effects on rating accuracy and rating errors. *Journal of Applied Psychology*, (114)12-42 1.
- California State Department of Education. (1993). *Science: New Dimensions in Assessment*. Sacramento, CA: California State Department of Education.
- Campbell, C.H., Campbell, R.C., Rumsey, M.G., & Edwards, D.C. (1986). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Campbell, C.H., Ford, P., Rumsey, M.G., Pulakos, E.D., Borman, W.C., Felker, D. B., de Vera, M. V., & Reigelhaupt, B. J. (1990). Development of multiple job performance measures in a representative sample of jobs. *Personnel Psychology*, 41277-300.
- Campbell, J.P. (Ed.). (1987). *Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1985 fiscal year*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

- Campbell, J.P, McHenry, Ji., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology*, 41313-334.
- Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin*, 51327-358.
- Gottfredson, L. (1990). The evaluation of alternative measures of job performance. In A.K. Wigdor & B.F. Green (Eds.), *Performance assessment for the workplace*, Vol. 2. Washington, DC: National Academy Press.
- Harris, D.A., McCloy, R.A., Dempsey, J.R., Roth, C., Sackett, P.R., Hedges, L.V., Smith, D.A., & Hopn, P.F. (199 1). Linking enlistment standards to job performance, Phase 1: A job performance model. Alexandria, VA: Human Resources Research Organization.
- McHenry, J.J., Hough, L.M., Toquan, LL, Hanson, M.A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and critical domains. *Personnel Psychology*, -42, 335-354.
- Pulakos, ED., & Borman, W.C. (1986). Development and field test of Army-wide rating scales and rater orientation and training program (ARI Technical Report 716). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Shavelson, R.J., Mayberry, P.W., Weichang, L., & Webb, N.M. (1990). Generalizability of job performance measurements: Marine Corps rifleman. *Military Psychology*, 2,, 129-144.
- Web, N.M., Shavelson, R.J., Kim, K.S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. *Military Psychology*, ,J, 91-110.
- Whetzel, D. L. (1991). Multidimensional screening: Comparison of a single-stage personnel selection/classification process with alternative strategies. Unpublished Doctoral Dissertation, George Washington University, Washington DC.
- Wigdor, A. K., & Green, B. F. (Eds.) (1991). *Performance Assessment for the Workplace*. Washington, DO National Academy Press.
- Wise, L.L, McHenry, J.J., & Campbell, J-P- (1990)- identifying optimal predictor composites and testing for generalizability across jobs and performance factors. *Personnel Psychology*, 41 355-366.
- Wise, LL. (1992). Lessons learned from military performance assessment. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wise, L.L., Peterson, N.G., Hoffman, R.G., Campbell, J.P., & Arabian, J.M. (1991). Army Synthetic Validity Project: Report of Phase III results. (ARI Technical Report 922). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Appendix A

Lessons Learned from Military Performance Assessment

(From Wise, 1992)

1. A careful specification of the domain to be assessed is critical.

In the military arena there was a long history of job analysis that provided a good basis for defining the measurement domain. The biggest debates were over whether measurement should reflect maximal performance (labelled proficiency by the NAS Committee) or typical performance and be broadened to include aspects of effectiveness such as teamwork or performance under adverse conditions. Results showing significant examinee-by-task interactions also suggest the need to assure appropriate coverage of the target domain.

In the educational arena, the relative success in defining the domain of mathematics for the state-by-state comparisons was surprising to many. Building a consensus for other subjects may be some-what more troublesome. The debate about whether and how to measure typical performance rather than just maximal performance is also very relevant. There is an explicit desire to create measures that reflect and hence motivate effort - an exam the student can and should study for.

2. An understanding of the uses to be made of the measures should precede their design.

The JPM project experienced tension between those who wanted to use the scores primarily to validate selection tests and those who also wanted to set performance standards that might be generalized across jobs. The former group was content with maximizing the reliable variability in individual differences. All they wanted was a good norm-referenced test. The latter group, wanted a carefully constructed, domain-referenced measure. The former group wanted an interval level scale for their correlational analyses. The latter group might have been satisfied with an ordinal scale for comparing individuals to standards, but actually argued for an absolute scale in order to simplify generalizations across jobs.

Multiple uses also have been discussed for improved educational achievement measures. If our purpose is to evaluate and improve educational systems, norm-referenced scales should suffice. If we want to address questions like how much education is enough (or how much should we spend on education), for either individuals or for the system as a whole, criterion-referenced scales may be more appropriate.

3. Attention should be paid to the type of scale that is required

The JPM performance scales had multiple uses. For some uses, such as dividing performance into acceptable versus unacceptable, an ordinal scale would suffice. For other purposes, such as placing dollar values on the utility of different performance levels, a ratio scale would be required. Item response theory (IRT) was developed in the educational arena. It provides a basis for generalizing scores across different sets of items or different samples of examinees. Many believe that it provides an "interval" scale of achievement. This is only technically true, within the framework of the measurement model. A different measurement model with monotonic transformations of the theta scale (and the associated item characteristic curves) could fit the data equally well. IRT models may not work well with more complex behavioral samples involving greater interdependencies among the scorable units. Also, a scale that suggests an "infinite" difference between knowing nothing and knowing a little may not be appropriate. The percent-GO, or percent-correct metric used in the JPM project provides a reasonable scale for competency assessment and should be considered in more complex educational assessment as well.

4. Written tests are not everything.

Knowing facts about a procedure and being able to execute it successfully are not exactly the same thing. The results varied considerably across and within jobs, but generally sufficient differences were found between written and hands-on performance tests to encourage educational researchers to press on in efforts to develop alternatives. So far, however, very little has been changed as a result of developing and using alternative measures.

5. Alternatives measures are to develop and even to administer and score.

This finding is not at all new to those who have been working on standardized writing assessments. Difficulties in developing equivalent prompts, in hiring and training scorers and the great time and costs associated with reading and scoring large numbers of essays are well known. In a time of great competition for educational dollars, the potentially modest benefits associated with better measures must be carefully weighed against the costs.

6. More attention might be given to analogs of performance ratings.

Tests are one-time events. A "portfolio" approach offers some promise for assessing typical rather than just maximal performance, but well-developed rating scales and well-designed rater

training programs may offer a lower cost alternative. Grades might be viewed as a form of rating, but frequently the desire for "objectivity" has removed any component of expert judgment. What is lacking is the type of standardization across teachers and schools that might be gained through carefully constructed and anchored scales. Other principles, such as the use of multiple raters and of rater training programs might also be integrated.

7. Procedures for assessing the generalizability of performance measures are important.

Results from the JPM effort indicated very significant examinee-by-task interactions. Military personnel researchers are still debating the extent to which performance measures can be generalized across jobs. This also is a key issue in educational assessment. More novel performance tasks may not generalize well outside a specific domain of knowledge.

¹ Note: The views expressed are those of the author and does not necessarily reflect the position of the Defense Manpower Data Center or other agencies within the Department of Defense.

Designing Scoring Rubrics For Performance Assessments: The Heart of the Matter

Judith Arter

Northwest Regional Educational Laboratory

Good quality performance criteria (often expressed in a scoring guide or rubric) should do more than merely give us a way to assign a score on a performance assessment. Good performance criteria can also help us define our instructional goals and targets for students, and can be used as an instructional tool in the classroom. In this paper I will argue that not only can performance criteria be used in this fashion, they must be designed to function in this fashion if we want our performance assessments to work.

There are several things that can diminish the usefulness of performance criteria as instructional tools. If we accept the premise that performance be designed to support instruction, then we need to build several features into them. These features will also be discussed in this paper.

Rationale For Considering Instructional Usefulness When Designing Performance Criteria

Reason # 1: Clearly stated performance criteria are excellent instructional tools

Would you agree with the following proposition:

Whenever we make a judgment about anything we use criteria whether we can articulate them or not.

Most people agree this is so. In classrooms, judgments about students are being made all the time. For example, teachers daily make judgments about students. These judgments are based on criteria. Therefore, there are only two choices: we can either make our criteria crystal clear to students or we can make them guess. How often have we made our students guess at criteria because we either have difficulty articulating them, were too busy to do so, or did not realize it was important to do so?

(Students have) been conditioned to believe that great papers 'just happen.' that they are a guessing game, and that one finds out what to do after it's too late (Krest, 1990).

Teachers, however, are not the only ones in the classroom making judgments. Not only do students make judgments about themselves and their work without encouragement, we are now systematically asking

students to self reflect and self-assess as a way for students to take control of their own learning, and as a way to accomplish some of the critical thinking goals we now hold for them. If all judgments are based on criteria, students use criteria during their self-assessments and self-reflections. We again have only two choices: either we can systematically assist students in exploring criteria that reflect the important dimensions of tasks (as articulated by themselves or others) or we can leave them on their own to struggle through as best they can.

I want (students) to see evaluation in its best sense - a source to inform teaching and learning. To that end we develop a vocabulary for commenting on the admirable and problematic aspects of writing. The more we examine samples, the richer and more helpful this language of evaluation becomes (Erickson, 1992).

"Winning points" may be the final goal of classroom work as it is of the sports endeavor, but the grade, like the final score of the game, never taught anyone how to win again, or why they lost. For the truly successful contenders, playing the game is always about learning the game ... however often it seems to be about scoring more wins than losses (Lucas, 1992).

The point here is that the development of good performance criteria is not just an exercise in developing an assessment tool that is external to the instructional process - one that is used only to "monitor" student progress. Rather, good performance criteria help teachers and students understand the targets of instruction: What is expected? What does good look like? What do I want to accomplish? How will I know when I'm there? What kind of feedback do I give to improve student work next time? Why did I win? again? They also provide a vocabulary for discussing work.

In fact, many teachers have reported that the process of developing the performance criteria is at least as useful as having the final assessments in place, because it forces them to articulate and come to agreement on what they value (Harman, 1992; Hebert, 1992; Murphy & Smith, 1990; Portfolio, 1990; Sugarman, 1989).

Reason #2: Who will be administering our performance assessments, anyway?

Classroom teachers will be, that's who. What will happen if they don't see the value or rationale in what they are doing? It might be that our wonderful, "authentic" performance assessments will not actually result in better information after all. If we, as large-scale assessors, don't take the needs of teachers into account, we also might not get what we want: better measures of student achievement.

Reason #3: Where does change occur?

Again, change occurs in the classroom. Assessment can only be an agent for change if it promotes change in the classroom. As seen above, clearly articulated performance criteria can change instruction. As long as we're spending so much money for performance assessments, why not build in features that increase the chance that they will be useful in the classroom as instructional tools?

What Do Instructionally Useful Performance Criteria Look Like?

If we accept the premise that performance assessments need to be designed with instructional usefulness in mind, we need to consider whether there are design features that are more or less effective in accomplishing this goal. There are, in fact, different types of performance criteria, and not all of them are designed with the instructional end-user in mind. In fact, the performance criteria designed for many assessments emphasize ease and efficiency over instructional usefulness. This is understandable for large-scale applications, but perhaps shortsighted.

Design Consideration #1: The Need For Generalized Criteria.

Consider the math performance assessment in Figure 1 from the publication *Riverside Curriculum Assessment System. Performance Test Exercises* (1991).

Note that the task is open-response (students construct a response). Also note that the criteria by which the performance is evaluated are tied directly to the task; that is, there are a different set of criteria for each task. To get a "4" on this task, the student has to draw a particular picture and provide a particular explanation. This type of scoring can be very efficient for large scale uses; it is both fast and reliable. The biggest problem is that it is not useful instructionally. It helps you to see what good performance looks like on this one task, but doesn't help you see what good math problem solving looks like in general. Scoring this exercise will not necessarily help you score the next one and will not help students be able to "win" next time.

I propose that we should be aiming for performance criteria that provide an overall picture of the target, not just how the target manifests itself in a single problem. For example, in the area of math problem solving, consider the four trait analytical model being developed by the state of Oregon, shown in Figure 2.

The same criteria are applied to all open-ended and open-response math problems. Evaluating one problem solution will help you evaluate the next solution because the goal is to understand what good

Performance Criteria Tied To Task

Task:

Mr. Ramirez helped four other students make key holders. He gave Rhonda, Sam, Tony, and Uta a board and told them to share it equally. first, Rhonda measured and cut one fourth of the board. next, Sam measured and cut one third of the remaining board. Finally, Tony measured and cut one half of the remaining board. Uta used the piece that was left. Did the four students hare the board equally? Draw a picture and explain your answer.

Performance Criteria:

A 4 response contains both a picture and an explanation that indicate a clear understanding of the pattern; contains a picture showing a whole divided into four equal parts, contains an explanation that enhances the picture by comparing the size of each piece using either sentences, computations, or a combination of both.

A 3 response contains a picture that indicates a clear understanding of the pattern but only an attempt at an explanation, contains an explanation that indicates a clear understanding of the pattern but only an attempt at a picture; has limited detail in the picture or the explanation.

A 2 response offers an adequate picture only; offers an adequate explanation only; contains an explanation that does not enhance the picture; may be difficult to understand due to errors in language and grammar.

A 1 response makes some attempt at a picture or an explanation, is unclear.

problem solving is, in general; the goal is generalization. Of course, in order to use the system well, raters (including teachers and students) have to see many samples of problem solutions. Thus, it takes longer to train raters than using systems where criteria are tied to tasks, and consistency of judgments between individuals, least at first, is lower. However, in the long run we, and students, will have greater understanding if we shoot for more generalized performance criteria. (Which would you rather have your students do: score a bunch of performance tasks where each task used different criteria, or a more generalized evaluation procedure that tried to analyze what makes problem solving good in general?)

Caveat time: When the assessment goal is measurement of conceptual knowledge, then specific ce criteria might be warranted. For example, consider the performance assessments developed by Lake County Educational Service Center (1992) in Illinois. They've developed a series of performance tasks to assess student ability to apply their knowledge of solid waste disposal. For example, students take a used lunch tray and indicate how each item might be reduced, reused, or recycled. Responses are scored degree of "correctness;" in other words, how well students understand the concepts involved.

When the goal is conceptual knowledge, it might be useful to have performance criteria that clearly

Figure 2

Oregon Four-trait Analytical Scoring Model for Math Problem Solving

Conceptual Understanding of the Problem: Conceptual Understanding includes the student's ability to interpret the problem and select appropriate information to apply a strategy for solution. Evidence of conceptual understanding is communicated through making connections between the problem situation, relevant information, and logical/masonable responses. Students demonstrate conceptual understanding in math when they provide evidence that they can use of interrelate models, diagrams, and varied representations of concepts; can compare, contrast, and integrate related concepts; and can interpret the assumptions and relations involving concepts in mathematical settings.

Procedural Knowledge: Procedural knowledge deals with the student's ability to demonstrate appropriate use of mathematics. Evidence of procedural knowledge is provided in the mathematics the student chooses to use and their ability to select and apply the appropriate procedures correctly. Procedural knowledge includes the various numerical algorithms in mathematics that have been created as tools to meet specific needs in an efficient manner. It encompasses the abilities to read and produce graphs and tables, execute geometric constructions, perform non-computational skills such as rounding and ordering, verify and justify the correctness of a procedure using concrete models or symbolic methods, and extend or modify procedures to deal with factors inherent in the problem setting.

Problem solving Skills and Strategies: Problem solving requires the use of many skills which are often used in certain combinations before the problem is solved. A combination or sequence of skills used in working toward the solution is referred to here as a strategy. Strong student responses should demonstrate the ability to use problem solving skills/strategies and demonstrate good reasoning that lead to a successful resolution of a problem.

Communication: Effective communication is essential to learning and knowing mathematics. Mathematics communication is demonstrated by the use of symbols and terms which attach specific, and sometimes different, meanings to common words. In assessing the student's ability to communicate mathematically, particular attention should be paid to both the meanings they attach to the concepts and procedures of mathematics and also their fluency in explaining, understanding, and evaluating the ideas expressed in mathematics.

articulate what conceptual understanding looks like for specific cases, rather than having a generalized scoring guide for conceptual knowledge, although I've seen performance criteria taking both approaches.

Design Consideration #2: Holistic v. Analytical Trait Systems

Both holistic and analytical trait systems require judgment. Holistic systems rate a performance as a whole -- one score. Analytical trait systems require judgmental ratings along several dimensions thought to be important. For example, consider Oregon's six-trait analytical scoring system for writing, shown in

Figure 3 and Appendix A. In this system, scores of 1-5 are given for each of the six traits of ideas, organization, voice, word choice, sentence fluency and conventions. (The same set of criteria are used for all types of writing.) It's not that holistic systems look for different features in the writing than analytical trait systems, it's just that they are all weighed together to arrive at the final, single score.

Holistic scoring is more efficient for large-scale assessment than analytical trait scoring. It is faster and raters can be trained more quickly. However, once again, holistic scores are not as useful instructionally. Two students can receive a "Y" for vastly different reasons; say one is very strong in conventions and weak in ideas and voice, while the other is strong in ideas and voice, but weak in conventions. This not only tends to be confusing to students, it also makes it more difficult to articulate to students what good writing looks like. Analytic trait systems communicate more specifically.

Training students to revise their writing trait by trait is a very powerful instructional tool. Consider the student work shown in Appendix B - two papers and a self-reflection. These papers were written by the same student, one at the beginning of the sixth grade, the other at the end during the first year the student's teacher taught the students to revise their writing trait by trait. The self-reflection clearly shows that the student understands why her writing has improved. (See Appendix B at the end of the article).

Assessments communicate and model what we value. Which is better to model in large-scale assessment, a quick holistic system, or an analytical trait system that has a great deal of instructional potential in the classroom? As long as we're doing performance assessment why not have it be an in-service tool as well?

Design Consideration #3: Covering All That Is Important

Good performance criteria cover the right "stuff." Consider the Informal Writing Inventory (Scholastic Testing Service, 1986), a screening instrument for special education. One portion of this assessment requires students to look at a picture and write a paragraph telling about what is happening in the picture. The paragraph is evaluated using the criteria shown in Figure 5.

Now, granted that this assessment might just focus on grammar, but still you would probably not want to draw any conclusions about student ability to write from just looking at grammar. Thus, the criteria by which this performance was evaluated do not cover all the relevant dimensions of the task.

Although this is an extreme case, some large-scale performance assessment systems tend in the same direction. For example, the Illinois analytical trait writing assessment does not score voice and word

choice because the developers felt that it would be too difficult to get consistency in scores; it is just too personal. What does this communicate to teachers about what is important to concentrate on in writing? What effect will this have on instruction? Anecdotal reports from teachers in Illinois indicate that, in fact, teachers do concentrate their instruction on the traits measured in the assessment. Do we leave things out just because they are difficult to define?

Design Consideration #4: Have teachers do the scoring.

Although this might not be a consideration for actually designing performance criteria, it is certainly a

Figure 5

Informal Writing Inventory Scoring Guide

The error index indicates the frequency with which errors are made ... In a 100-word passage, if 60 errors were produced, the error index could be expressed as 60/100, or 60% The communication index is the ratio of errors that disrupt communication to total errors ... As the communication index approaches 12, the likelihood increases that the writer is disabled. However, the error index and the communication index can be interpreted and validated only by reference to each other. Whereas the error index indicates the number of mistakes, the communication index indicates the writing quality." (Informal Writing Inventory, 1986, pp. 4-5).

Three types of errors shown in the scoring guide include: incorrect abbreviations, misspellings, poor punctuation, incorrect capitalizations, incorrect grammar, illegible writing, sentence fragments, and incorrect use of plurals (p. 9). These are grouped into the three areas of handwriting, spelling, and grammar (p. 15).

design consideration for the performance assessment in general. Why spend all that money to have someone else learn what you value in a performance? If teachers are doing the scoring, they are learning information that can be taken back into the classroom to improve instruction. (Assuming, of course, that the performance criteria are designed according to points #1-3 above.) What about "objectivity?" Our experience is that teachers, just like everyone else, can be trained to be consistent in scoring.

Conclusion

If we want to maximize the impact of our performance assessment dollar, we should design performance criteria that teachers can use in the classroom. If we don't, our "new" assessments might not have any more impact than our "old" assessments.

References

- Erickson, M. (1992). Developing student confidence to evaluation writing. *Quarterly of the National Writing Project & The Center For The Study of Writing and Literacy*, 11 7-9.
- Giordano, G. (1986). *Informal Writing Inventory*. Bensenville, IL: Scholastic Testing Service, Inc.
- Harman, S. (1992). The basal 'conspiracy.' In K. Goodman, L. Bird, and Y. Goodman (Eds.), *The Whole Language Catalog Supplement on Authentic Assessment*. Santa Rosa, CA: American School Publishers.
- Hebert, E.A. (1992). Portfolios invite reflection - from students and staff. *Educational Leadership*, 4E 58-61.
- Krest, M. (1990). Adapting the portfolio to meet student needs. *English Journal*, a 29-34.
- Lake County Educational Service Center. (1992). *Discovering the problem of solid waste: Performance assessments*. Lake County Educational Service Center, 19525 W. Washington St., Grayslake, IL 60030.
- Lucas, C. (1990). Introduction: Writing portfolios - changes and challenges. In K. Yancey (Ed.), *Portfolios in the writing classroom*. Urbana, EL: National Council of Teachers of English.
- Murphy, S. and Smith, M. (1990). Talking about portfolios. *The Quarterly of the National Writing Project & The Center For The Study of Writing and Literacy*. Spring, 1990, 1-3 24-27.
- Portfolio. (1990). The evolution of PROPEL? *Portfolio: The Newsletter of Arts PROPEL*, 1, p. 1.
- Riverside Publishing Company. (1991). *The Riverside Curriculum Assessment System: Performance Test Exercises*. Chicago, IL: Riverside Publishing Company.
- Sugarman, J. (1989). Teacher portfolios inform assessment. *Educator*. May 1989, 5-6.

Appendix A

Oregon's Six Trait Analytical Scoring System

Ideas and Content (Development)

- 5: This paper is clear, focused, and interesting. It holds the reader's attention. Relevant anecdotes and details enrich the central theme or story line. Ideas are fresh and original.**

The writer seems to be writing from experiences and shows insight: a good sense of how events unfold, how people respond to life and to each other.

Supporting, relevant, telling details give the reader important information that he or she could not personally bring to the text.

The writing has balance: Main ideas stand out

The writer seems in control and develops the topic in an enlightening, entertaining way.

The writer works with and shapes ideas, making connections and sharing insights.

- 3: The paper is clear and focused. The topic shows promise, even though development is still limited, sketchy or general.**

The writer is beginning to define the topic, but is not there yet. It is pretty easy to see where the writer is headed, though more information is needed to "fill in the blanks."

The writer does seem to be writing from experience, but has some trouble going from general observations to specifics.

Ideas are reasonably clear and purposeful, even though they may not be explicit, detailed, personalized, or expanded to show in-depth understanding.

Support is attempted, but doesn't go far enough yet in expanding, clarifying, or adding new insights.

Themes or main points seem a blend of the original and the predictable.

- 1: As yet, the paper has no clear sense of purpose or central theme. To extract meaning from the text, the reader must make inferences based on sketchy details. More than one of the following problems is likely to be evident:**

Information is very limited or unclear.

The text is very repetitious, or reads like a collection of random thoughts from which no central theme emerges.

Everything seems as important as everything else; the reader has a hard time sifting out what's critical.

The writer has not yet begun to define the topic in a meaningful or personal way.

The writer may still be in search of a real topic, or sense of direction to guide development.

Organization

5: The organization enhances and showcases the central idea or theme. The order, structure, or presentation is compelling and moves the reader through the text.

Details seem to fit where they're placed; sequencing is logical and effective.

An inviting introduction draws the reader in and a satisfying conclusion leaves the reader with a sense of resolution.

Pacing is very well controlled; the writer delivers needed information at just the right moment, then moves on.

Transitions are smooth and weave the separate threads of meaning into one cohesive whole.

Organization flows so smoothly the reader hardly thinks about it.

3: The organizational structure is strong enough to move the reader from point to point without undue confusion.

The paper has a recognizable introduction and conclusion. The introduction may not create a strong sense of anticipation; the conclusion may not leave the reader with a satisfying sense of resolution.

Sequencing is usually logical. It may sometimes be too obvious, or otherwise ineffective.

Pacing is fairly well controlled, though the writer sometimes spurts ahead too quickly or spends too much time on the obvious.

Transitions often work well; at times though, connections between ideas are fuzzy or call for inferences.

Despite a few problems, the organization does not seriously get in the way of the main point or storyline.

1: The writing lacks a clear sense of direction. Ideas, details or events seem strung together in a random, haphazard fashion--or else there is no identifiable internal structure at all. More than one of the following problems is likely to be evident:

The writer has not yet drafted a real lead or conclusion

Transitions are not yet clearly defined; connections between ideas seem confusing or incomplete.

Sequencing, if it exists, needs work.

Pacing feels awkward, with lots of time spent on minor details or big, hard-to follow leaps from point to point.

Lack of organization makes it hard for the reader to get a grip on the main point or storyline.

Voice

5: The writer speaks directly to the reader in a way that is individualistic, expressive, and engaging. Clearly, the writer is involved in the text and is writing to be read.

The paper is honest and written from the heart. It has the ring of conviction.

The language is natural yet provocative; it brings the topic to life.

The reader feels a strong sense of interaction with the writer and senses the person behind the words.

The projected tone and voice give flavor to the writer's message and seem very appropriate for the purpose and audience.

3: The writer seems sincere, but not genuinely engaged, committed, or involved. The result is pleasant and sometimes even personable, but short of compelling.

The writing communicates in an earnest, pleasing manner. Moments here and there amuse, surprise, delight or move the reader.

Voice may emerge strongly on occasion, then retreat behind general, vague, tentative, or abstract language.

The writing hides as much of the writer as it reveals.

The writer seems aware of an audience, but often to weigh words carefully, to stand at a distance, and to avoid risk.

1. The writer seems indifferent, uninvolved or distanced from the topic and/or the audience. As a result, the writing is flat, lifeless or mechanical; depending on the topic, it may be overly technical or jargonistic. More than one or the following problem is likely to be evident:

The reader has a hard time sensing the writer behind the words. The writer does not seem to reach out to an audience, or make use of voice to connect with that audience.

The writer speaks in a kind of monotone that tends to flatten all potential high's and low's of the message.

The writing communicates on a functional level, with no apparent attempt to move or involve the reader.

The writer is not yet sufficiently engaged or at home with the topic to take risks or share him-/herself.

Word Choice

5: **Words convey the intended message in an interesting, precise, and natural way. The writing is full and rich, yet concise.**

Words are specific and accurate; they seem just right.

Imagery is strong.

Powerful verbs give the writing energy.

Striking words and phrases often catch the readers eye, but the language is natural and never overdone.

Expression is fresh and appealing; slang is used sparingly.

3: **The language is functional, even if it lacks poach; it does get the message across.**

Words are almost always correct and adequate (though not necessarily precise); it is easy to understand what the writer means.

Familiar words and phrases communicate, but rarely capture the reader's imagination. The writer seems reluctant to stretch.

The writer usually avoids experimenting-, however, the paper may have one or two fine moments.

Attempts at colorful language often come close to the mark, but may seem overdone, or out of place.

A few energetic verbs liven things up now and then; the reader yearns for more.

The writer may lean a little on redundancy, or slip in a cliches-but never relies on thew crutches to the point of annoyance.

1: **The writer struggles with a limited vocab ulary, searching for words to convey meaning. More than one of the following problems is likely to be evident:**

Language is so vague and abstract (e.g., *It was a fun time, it was nice and stuff*) that only the most general message comes through.

Persistent redundancy clouds the message and distracts the reader.

Cliches or jargon serves as a crutch.

Words are used incorrectly in more than one or two cases, sometimes making the message hard to decipher.

The writer is not yet selecting words that would help the reader to a better understanding.

Sentence Fluency

5: The writing has an easy flow and rhythm when read aloud. Sentences are well built, with consistently strong and varied structure that makes expressive oral reading easy and enjoyable.

Sentence structure reflects logic and sense, helping to show how ideas relate. Purposeful sentence beginnings guide the reader readily from one sentence to another.

The writing sounds natural and fluent-, it glides along with one sentence flowing effortlessly into the next.

Sentences display an effective combination of power and grace.

Variation in sentence structure and length adds interest to the text.

Fragments, if used at all, work well.

Dialogue, if used, sounds natural.

3. The text hums along efficiently for the most part, though it may lack a certain rhythm or grace. It tends to be more pleasant or businesslike than musical, more mechanical than fluid.

The writer shows good control over simple sentence structure, more variable control over complex sentence structure.

Sentences may not seem skillfully crafted or musical, but they are grammatical and solid. They hang together. They get the job done.

The writer may tend to favor a particular pattern (e.g., subject-verb, subject verb), but there is at least some variation in sentence length and structure (Sentence beginnings are NOT all alike).

The reader sometimes has to hunt for clues (e.g., connecting words like *however therefore, naturally, on the other hand, to be specific, for example, next, first of all, later, still, etc.*) that show how one sentence leads into the next.

Some parts of the text invite expressive oral reading; others may be a little stiff, choppy or awkward. Overall, though, it's pretty easy to read this paper aloud if you practice.

1: The paper is difficult to follow or read aloud. Most sentences tend to be choppy, incomplete, rambling, or awkward; they need work. More than one of the following problems is likely to be evident:

Sentences do not sound natural, the way someone might speak. Word patterns are often jarring or irregular, forcing the reader to pause or read over.

Sentence structure tends to obscure meaning, rather than showing the reader how ideas relate.

Word patterns are very monotonous (eg., subject-verb, subject-verb-object). There is little or no real variety in length or structure.

Sentences may be very choppy. Or, words may run together in one giant "sentence linked by "ands 's" or other connectives.

The text does not invite expressive oral reading.

Conventions

- 5: The writer demonstrates a good grasp of standard writing conventions (e.g, grammar, capitalization, punctuation, usage, spelling, paragraphing) and uses them effectively to enhance readability. Errors tend to be so few and minor the reader can easily skim right over them unless specifically searching for them.**

Paragraphing tends to be sound and to reinforce the organizational structure.

Grammar and usage are correct and contribute to clarity and style.

Punctuation is smooth and guides the reader through the text.

Spelling is generally correct, even on more difficult words.

The writer may manipulate conventions-particularly grammar-for stylistic effect.

The writing is sufficiently long and complex to allow the writer to show skill in using a wide range of conventions (This criterion applies at grade 7 and up only)

Only light editing would be required to polish the text for publication.

- 3: The writer shows reasonable control over a limited range of standard writing conventions. However, the paper would require moderate editing prior to publication. Errors are numerous or serious enough to be somewhat distracting, but the writer also handles some conventions well.**

Spelling is usually correct (or reasonably phonetic) on common words.

Terminal (end-of-sentence) punctuation is almost always correct; internal punctuation (commas, apostrophes, semicolons) may be incorrect or missing.

Problems with grammar or usage are not serious enough to distort meaning.

Paragraphing is attempted. Paragraphs sometimes run together or begin in the wrong places.

The paper seems to reflect light, but not extensive or thorough, editing.

- 1: Errors in spelling, punctuation, usage and grammar, capitalization and/or paragraphing repeatedly distract the reader and make the text difficult to read. More than one of the following problems is likely to be evident:**

The reader must read once to decode, then again for meaning.

Spelling errors are frequent, even on common words.

Punctuation (including terminal punctuation) is often missing or incorrect.

Paragraphing is missing, irregular, or so frequent (e.g., every sentence) that it does not relate to organization of the text.

Errors in grammar and usage are very noticeable, and may affect meaning.

Extensive editing would be required to polish the text for publication.

Appendix B

A Sixth Grade Student's Writing and Self-Reflection

California

(written at the beginning of sixth grade)

I went down to my Grandma's house in California and I got to ride horses, I swam, went shopping, saw old friend, I went to a party, I went to the new Marine World, and I had a lot of fun.

I drove down there in a car with my uncle and drove back with him. He went down to visit his mom and dad so it worked out pretty good.

I used to live in California till a year ago. It was a 30 min. drive away from San Francisco I live in Walnut Creek. I went to school at Walnut Acres for 4 years, since 2nd grade. Then moved to Oregon and we bought a gas station. When I got to go I was glad and happy my mom let me. I HAD FUN!

A Little Mouse Statue

(written at the end of sixth grade)

Every time I walk in my room, or pass my dresser, I see something that's very special to me. It is a little statue of a mouse. His tiny hands are expanded as far apart as they allow themselves to be. And, at the bottom of the statue it reads, "I love you this much."

I believe I was four years old when my grandma took me over to her bedroom closet one day and got my statue off the very top shelf. Then, with extreme care, she unwrapped a small object and handed it to me. It was the mouse statue.

Ever since then, even now, I have him placed on my dresser to admire every time I pass my dresser, or stand next to my dresser dressing or putting on earrings, I think of my grandma.

I think of the way my grandma always expanded her arms and said, "I LOVE YOU THIS MUCH" just like the little mouse statue does. And, I'd do the same. Then we'd hug each other followed by enormous kisses. Her gentle and kind smile, the glitter in her eyes, and the way she always stuck up for me if I was in a fight with my mom are all things I remember about her. Today, she still takes me special places, and she's always there if I need someone to talk to or get advice from.

I will always treat my statue with the most respect just like my grandma asked me to. And I will

always treasure its unique way of making me feel close to my grandma even when she's not around every time I glance at him. And who knows, maybe one day I'll be giving him to my granddaughter!!!

Student's Self-reflection

(written at the end of sixth grade)

I have become a better writer this year. I have learned to put more focus in my writing and stick with one topic. I think about my topic before I write, and I share my writing in a writing group. That is something I did not like to do at first, but now I do. I think my writing has a lot more voice now. Voice is the part of your writing that shows how you feel about your topic because the thoughts and feelings come from your heart. This year we read Charlotte's Web, and that is a book that I think has a lot of voice. I have also worked very hard on my word choice. I try to find just the right word to say what I mean and not just the first word that comes into my mind. The way I have grown the most is that I like to write a lot more than I used to, especially poems. I think I could be a poet if I wanted to, and I think my writing shows that.

Discussant's Comments

Joe McDonald

Brown University

I would like to acknowledge that I appreciated all these papers, and learned a lot from reading across them. Nonetheless, I am willing to play the part of critical friend.

First, I would substitute the word *intersubjectivity* for *objectivity* in the title of the symposium. I know that gives us an impossible tide -- "intersubjectivizing the subjective," but I'll do anything to make a point. And the point is not just verbal - arguing about word choice - or philosophical - trying to replace a positivist outlook with a phenomenological one. It's really a practical point. I'm interested in building new accountability systems that deepen the education of all kids. I'm also painfully aware that accountability systems do not usually do this, and that at least part of the reason why they do not has to do with their pursuit of objectivity.

Now I know that it is not obvious how to arrange things otherwise. That is especially why I appreciate the efforts of these authors. Their work, like the work of many others today in the communities of practice, measurement, and policy, help us to conceive of accountability in some other terms other than those associated with the duality of subjective and objective.

The Perlman Paper

There is good news and bad news in the Perlman paper. The good news is that roughly 45% of survey respondents have developed performance assessments and that the pattern of their responses shows awareness of the power of performance assessment construction as a professional development activity.

The bad news is that the survey suggests this power is currently in use at the district and classroom levels, but not at the school level (where it might enrich accountability in the fullest sense), and in writing but not in other subjects.