

## **Symposium 1**

### **Legitimate Ways To Prepare Students For Standardized Tests**

The current American phenomenon of attempting to prove the worth of most educational matters on the basis of standardized test scores has placed a spotlight on the testing environment. An outgrowth of NATD's 1985 symposium on cheating was a study of district practices in preparing students for tests.

Glynn Ligon moderated a most fascinating symposium -- unique in the degree to which audience participation was encouraged and did, in fact, occur. Using a survey of school district practices by David Covert as the focal point, panelists and audience entered into a lively discussion of practices and concerns.

Kevin Matter shared some principles which he follows towards assuring the legitimacy of preparation. Ellen Pechman argued that of the same import as the amount of preparation is a focus on equity in preparation. Peter Wolmut proposed that problems in this area will continue as long as we have expectations of teacher test administration behavior, but neither give them clear guidelines nor advise them of the consequences of that behavior. Brenda Loyd brought the discussion full circle, pointing out that a great deal of energy is being expended on matters for which testing was never intended.

# **Test Preparation: An Overview**

**Glynn Ligon**

## **Austin TX Independent School District**

I have long had strong feelings about test preparation activities. Nothing undermines the validity of an achievement testing program more than inequities and cheating in preparing students for testing. Even the critical issue of test security is tied to ensuring only legitimate preparation activities.

A brief summary of our approach is:

- Plan ahead to establish clear guidelines for teachers to follow.
- Ensure test security.
- Provide practice tests and exercises to all teachers.
- Make test preparation a year-long activity by incorporating some characteristics of standardized tests into teacher-made tests.
- Follow up on any questionable activity.

Three AERA papers related to legitimate ways to prepare students for standardized testing describe in greater detail our district's philosophy and procedures for ensuring consistent and legitimate test preparation activities:

M. Kevin Matter & Glynn Ligon, "Achievement Test Preparation: A Year-Long Goal, Not a Last Minute Thought" (AERA, Montreal; 1983)

M. Kevin Matter.& Glynn Ligon, "Preparing Students for Standardized Testing: Everybody's Business" (AERA, New Orleans; 1984)

Glynn Ligon, "Opportunity Knocked Out: Reducing Cheating by Teachers on Student'.7psts" (AERA, Chicago; 1985)

A fourth publication provides a more comprehensive overview of our system:

Glynn Ligon, "Preparing Students for Standardized Testing" in W.E. Hathaway (ed.), Testing in the Schools, New Directions for Testing and Measurement, No. 19 (San Francisco: Jossey-Bass, 1983)

## The NATD Test Preparation Survey

R. David Covert

### Jefferson County KY Public Schools

In August, 1985 a proposal for a symposium was submitted to the National Council on Measurement in Education by Paul Brown from Indianapolis on "Legitimate Ways To Prepare Students for Testing".

Standardized tests are now widely used to judge educational success or failure of states, districts schools and contiguous systems or districts. Because of the accountability "pressure" this creates, there have been several incidents of undesirable test practices, i.e., teaching the test, found nationally. Yet many districts provide their students with useful, legitimate test-wisness activities. This survey was an attempt to determine the practices and policies which were operating nationally, at least in the school districts of NATD members.

The survey questions were created through a cooperative effort of R. David Covert/Jefferson County Schools, Louisville and Dr. Ed. Binkley/Metropolitan Nashville (Tennessee) Schools. The questions were created to address the issues to be raised in the symposium, from literature surrounding test taking practices, from questions raised by newspaper articles in cities where test taking practices were questioned, and from current trends of districts and states having test ethics codes.

The survey was designed with questions with a "quick" yes/no response and also provided space for an expanded answer, if desired. This was done since research into test taking skills found that there was little congruence among the NATD member districts on this issue. Practices varied widely from state to state/district to district.

The survey was sent to all NATD members who appeared on the Membership Directory list and its addendum for 1984-85. Of the 148 active members listed, 70 members, or 47% responded to the survey in time to have the data submitted. An additional six (6) surveys were received after the data was submitted for use at the NCME meeting, bringing the total responses to nearly 50%.

In addition to the objective responses and additional comments for each question, many members sent copies of manuals and materials used by their districts in student test preparation.

# NATIONAL ASSOCIATION OF TEST DIRECTORS

## TEST PREPARATION SURVEY

### RESULTS

APRIL, 1986

SAMPLE SIZE = 52 (100%).

Number of different states represented = 24 (3 unidentified).

States with the largest returns:

- Florida = 6 (11.5)
- Texas = 5 (9.6)
- California = 4 (7.7)

Others include:

- Wisconsin and Maryland each had 3 respondents.
- Michigan, Arizona, Massachusetts, Oregon, Tennessee, Nevada, Oklahoma, Ohio, and Indiana each had 2 respondents.
- Utah, Connecticut, Alabama, Kentucky, Virginia, Colorado, Georgia, Nebraska, Louisiana, and Illinois each had 1 respondent.

Respondents enclosing material on testing procedures:

Yes = 19 (36.5)

No = 33 (63.5)

1. Does your district provide assistance to the local school in identifying acceptable standardized test preparation activities?

Yes = 45 (86.5)

Guidelines are (N = 45):

Formal = 20 (44.4)

Informal = 25 (55.6)

Applicable to all tests = 28 (62.2)

Specific to district tests = 13 (28.8)

No = 7 (13.5)

2. Do any of your schools use practice tests (other than those which may be provided by the publishers) for classroom use for the standardized tests?

Yes = 37 (71.2)

Districtwide use? (N = 37)

Yes = 13 (35.1)

No = 23 (62.2)

Missing = 1 (2.7)

No = 15 (28.8)

Missing = 0

3. Do your schools have practice activities for the standardized test your district uses?

Yes = 34 (65.4)

Districtwide use? (N = 34)

Yes = 14 (41.2)

No = 17 (50.0)

Missing = 3 (8.8)

No = 16 (30.8)

Missing = 2 (3.8)

4. Do some teachers in your district present general content instruction prior to administration of a systemwide test?

Yes 26 (50.0)

Districtwide use? (N = 26)

Yes = 7 (26.9)

No = 19 (73.1)

No = 19 (36.5)

Missing = 7 (13.5)\*

\*NOTE: Six persons wrote in that they were confused by the term "general content instruction".

Are there any stipulations on the conducting of these activities?

Yes = 25 (48.1)

No = 9 (17.3)

Missing = 18 (34.6)

5. Does your state have a formal test Code of Ethics?

Yes = 4 (7.7) [Alabama and Kentucky]

No = 33 (63.5)

Missing = 4 (7.7)

Don't Know = 11 (21.2)

6. Does your district have a formal test Code of Ethics?

Yes = 8 (15.4)

No = 41 (78.8)

Missing = 3 (5.8)

7. Does your system have a formal (written) test security procedure?

Yes = 24 (46.2)

No = 27 (51.9)

Missing = 1 (1.9)

8. Six (11.5) respondents indicated either some media coverage or incident with teachers that lead them to review their testing preparation procedures. The consequences, if they were described, were to discipline the individual responsible and formalize a written testing procedure.

<b>District</b>	<b>Contact Person</b>
Akron, Ohio	John A. Stewart, Superintendent
Polk County, Florida	Director, Department of Testing 2051 Senate Street New Orleans, Louisiana
New Orleans, Louisiana	J. Jolly, Program Evaluation Building 502 3323 Belvedere Road West Palm Beach, Florida 33402
Palm Beach County, Florida	Dr. Perlman Department of Research and Evaluation 1819 W. Perhsing Rd. Chicago, Illinois 60609
Chicago, Illinois	Dr. Evangelina Mangino Office of Research and Evaluation 6100 Gaudalupe Street Austin, Texas 78752-4495
Austin, Texas	Dr. Evangelina Mangino Office of Research and Evaluation 6100 Guadalupe Street Austin, Texas 78752-4495

Some Districts With Test Preparation and Administration Guidelines

<b>School District</b>	<b>Contact Person</b>	<b>Publication Title</b>
New Orleans	Director, Department of Testing 2051 Senate Street New Orleans, LA	School Test Coordinator's Manual
Polk County, Florida	John A. Stewart Superintendent	Test Administration Handbook
Palm Beach, Florida	J. Jolly, Program Evaluation	Guidelines for Conducting Standardized Testing Programs
Chicago, Illinois	Dr. Perlman Dept. of Research and Evaluation 1819 W. Pershing Road Chicago, Illinois 60609	Test-Wiseness strategies to be used on the California Achievement Tests
Akron, Ohio	James Hardy, Assistant Superintendent of Curriculum and Instruction	Activities, Tips, and Clues to Help Make Test-taking Easy
Austin, Texas	Evangelina Mangino, Office of Research and Evaluation Austin, Texas 78752-4495	Guidelines for Test Administrators #80.64
Huntsville, Alabama	Martha Beckett, Office of the Superintendent P.O. Box 1256 Huntsville, Alabama 35807-4801	Test Security Guidelines
Dallas, Texas	Linus D. Wright, Superintendent 3700 Ross Ave. Dallas, TX 75204	Guidelines for Ensuring Valid Test Data
Mesa, Arizona	Dr. James DeGracie 549 N. Stapley Mesa, Arizona 85203	Directives for Student Testing
Milwaukee, Wisconsin	Department of Educational Research and Program Assessment 5225 W. Vliet Street Milwaukee, WI 53208	Guidelines for City-wide Standardized Testing
Tulsa, Oklahoma	Stan Harrison, Instructional Research and Evaluation P.O. Box 470208 Tulsa, OK 74147-0208	Approved Test Security and Administration Procedures

Cincinnati, Ohio	Bernard M Barbadora Testing Section 230 E. 9 <sup>th</sup> Street Cincinnati, OH 45202	Principal's/Test coordinator's Handbook for the California Achievement Tests
Oklahoma City, Oklahoma	Patricia J. Watson Planning and Research 900 N. Klein Oklahoma City, OK 73106	CAT Cookbook: Detailed Instructions for Administering the California Achievement Test
Jefferson County, Kentucky	R. David Covert Research Department 4409 Preston Highway Louisville, KY 40213	School-Based Certification Book
Memphis, Tennessee	Dr. Kathy Pruett, Division of Research Services, Room 114 Memphis City Schools 2597 Avery Avenue Memphis, Tennessee 38112	Test-Taking skills: A Guide For Teachers
Springfield, Massachusetts	Dr. John H. Howell Director of Research 195 State Street Springfield, MA 01103	Testing: One, Two, Three
Portland, Oregon	Peter Wolmut Multnomah Educational Service District 220 S.E. 102 <sup>nd</sup> St. Portland, OR 97216	Testing Bulletin

Legitimate Ways To Prepare Students For Testing:  
Being Up Front To Protect Your Behind

M. Kevin Matter  
Cherry Creek CO Schools

Note: The views expressed here are solely the Author's and do not represent the perspectives or policies of the organization with which he is affiliated.

The results of the National Association of Test Directors (NATD) test Preparation survey provide evidence that most of us are somewhat lax in providing district personnel with proper and sufficient information, procedures, and practices on preparing students for standardized testing. While many provide assistance in identifying acceptable test preparation activities and provide practice tests, few districts/states have a formal test Code of Ethics or written test security procedures.

But when does "some" information become "too much", and how many procedures are sufficient to maintain test security and not violate the tests' norms, yet provide teachers flexibility and freedom in instruction? These are difficult questions for any district to adequately answer.

An overall perspective that may help in developing procedures can be summarized as "being up front to protect your behind"

1. Communicate your agenda. Inform teachers, administrators, and parents about the importance of test security (e.g., the cost of booklets or adopting a new test; assuring usefulness of test results) and other test-related issues. This needs to be an ongoing effort, communicated not only in written form, but in presentations to various groups and in your own practices.
2. Determine / reaffirm the district's priorities. While we know a test score provides one indicator of performance taken at one point in time on one particular instrument, we cannot inhibit others from overemphasizing the importance of one score. We can, however, ensure that our efforts to prepare students for testing support effective instructional methods.

Ideally, test preparation activities should not be additional activities imposed upon teachers. Rather, they should be incorporated into the regular, ongoing instructional activities whenever possible. Particularly at the elementary school level, we need to protect instructional time and must ask ourselves if the planned test preparation activity will improve performance more than the same amount of instruction. Several publishers are in the process of developing extensive test preparation packages for specific standardized achievement tests. Materials such as these should be thoroughly examined before their use to determine if the techniques and methods presented are "instructionally" relevant and useful, beyond their test-preparation focus.

3. Determine the user's real need. Following point number one, it is important to determine the user's real information need. For example, a principal may request a test booklet for examination because a teacher wants more information about the test. While a test booklet may meet that need, other materials may provide more useful information. If test format is the "real" concern, a teacher's guide outlining the skills assessed would provide much more useful information than the test booklet itself.
4. Educate when possible. Many test preparation (particularly test-wiseness) activities and information are not relevant to the majority of items on standardized achievement tests. Techniques which may more effectively prepare a student for optimal performance on a standardized achievement test may not be appropriate for a teacher-made classroom examination. Likewise, several test-wiseness techniques may help a student achieve a higher score on a "pop quiz," but have no practical utility on a standardized test constructed by an expert test-item writer. Test preparation activities should enhance a teacher's ability to assess student progress on regular classroom work. These activities should help teachers write better test items which measure each objective level, using a variety of item formats and evaluation techniques. Test preparation activities should assist teachers in developing their own expertise in testing.
5. Spread the responsibility appropriately. Preparing students for standardized achievement testing is not a one person or a one office responsibility. Obviously, the best preparation is effective and thorough instruction on the objectives deemed important to learn. However, everyone in the school community has a responsibility and a role to play, including teachers, students, administrators, parents, and the public at large. Each group needs to be informed of their role and involved in carrying out their responsibilities in ensuring a valid test administration and in effectively (and correctly) using the results to improve instruction.

In summary, most of us need to more thoroughly provide materials, guidelines, and a philosophy of preparing students for standardized achievement testing. An open, honest, collaborative approach with district personnel and the community may be useful in their development.

Legitimacy is in the Eye of the Stakeholder

Ellen Pechman

Public School Forum of North Carolina

Until the onset of standardized promotional and graduation tests of "basic skills," only the most competitive students planning to enter the top colleges and professions were interested in becoming better prepared for tests. However, test preparation has recently taken a new turn. Improving SAT scores for better college placement is no longer the "hot issue." Instead, scoring optimally on basic skills and competency tests has become a matter of school and community pride or, at the individual level, it is the required ticket to grade and school promotion.

The changes in the goals of district-wide testing programs have also shifted the emphasis away from the individual student. In recent years, the impetus for developing district-wide test preparation comes not from students, but from numerous self-interested constituents: superintendents whose employment rests on the district's achievement levels, school boards whose election depends on demonstrating district-wide achievement, principals looking toward promotion, teachers and teacher unions concerned because students' achievement is part of the employee evaluation process, and parents seeking special education programs or worried that their youngsters might not be promoted. Also, in the past decade, courts have become an increasingly active constituent group in their use of test scores to monitor and evaluate whether districts are providing equitable instructional opportunities.

This shift in who cares about test scores has also affected what is considered "appropriate" preparation for tests. A look back into the closets of counseling and pupil personnel offices in school districts (those that predated our present testing and research departments) will uncover faded booklets carefully articulating test administration guidelines and suggestions about how best to prepare students so their test scores represent optimal performance levels. Tests and measurement books continue to include chapters on what kind of test preparation best readies students to achieve at their highest levels (e.g., Gronlund, 1985; Cronbach, 1970; Nitko, 1983; Thorndik6 and Hagen, 1977). In addition, ".tests and measurement" courses were, until recent years, a fundamental requirement in most teacher preparatory programs. In these courses, teachers were carefully instructed in acceptable practices of test development, administration, and student preparation.

\*Ellen Pechman is the former Director of the Department of Districtwide Testing, New Orleans Public Schools.

The literature detailing legitimate test preparation and coaching builds on those earlier foundations. Moreover, school districts large and small design and distribute test preparation instructions that are comprehensive and sensitively developed, reflecting much of the established wisdom on test administration procedures and the value and limits of prior test-focused preparation. These directions, Combined with the results of recent research on test preparation (see Gronlund, 1985 and Nitko, 1983 for good summaries) suggest approaches for preparing students for standardized tests consistent with those prescribed in the early days of testing: a solid academic foundation, knowledge of the test format and context, and comfort in the test situation.

Test preparation has been with us for a long time, and what is legitimate is well established. In my opinion, the new questions about legitimacy are more related to equity than to test issues. Families with high aspirations for their children have always "prepared" them for the tests they will encounter in life, coaching them at the swimming pool, around the dinner table, and on long car trips; teaching them to be comfortable with every test situation; and helping them develop skills in focusing, guessing, and responding flexibly to new intellectual challenges. Those who could afford it often also sent their children to special schools or tutors to have them instructed in these techniques. This instruction included much that now might be labeled as "teaching the test," for example, focused vocabulary drills and math problems based on items used in special editions or older versions of actual tests.

Promotional and graduation testing directs test preparation to a new group of students who, heretofore, have been educationally disenfranchised. In numerous ways, school achievement has eluded these students and, not surprisingly, their weak preparation is demonstrated by their consistently poor test performance. Thus, today's test preparation procedures attempt to remedy the failure to meet these students' needs in the classroom by extending the educational privileges that once belonged only to a few students. When done well, this is accomplished by providing basic academic content, knowledge of test format and context, and comfort in the test situation through the use of test-like items or items from previous test editions. But now there is a difference. Instead of approval, we encounter the suspicion commonly observed when we open a door to the educational underclass. Coaching for test success, a natural part of growing up for the luckiest among us causes alarm and cries of "teaching the test" when it is shared with those previously left out.

For the "new" achievement-conscious test taker, what, then, is legitimate test preparation? School district testing offices have found themselves informally experimenting with various test preparation procedures; honing those that raise district-wide scores in ways that achieve widespread acceptance; and dropping those that have little effect, raise controversial questions, or lead to erratic test scores. For example, most test-centered school districts mandate that teachers integrate test-taking skills and the district's curricular objectives into their instructional routine throughout the year. In some places, teachers are directed -- formally or informally -- to teach certain question formats and to concentrate their instruction primarily on

teaching the district's "testable" objectives. It is widespread practice in some environments to teach all the vocabulary words from tests that might cause difficulty; other districts have received severe criticism for incorporating the specific vocabulary to be found in the tests into the curriculum.

In such instances, are we teaching the curriculum and the students, or are we teaching the tests? Or, are we doing both? If achieving adequate scores on basic skills tests is the goal, such test preparation may be the most direct means of accomplishing it. In my view, this is both a philosophical and a pedagogical debate. The philosophical issues regarding what is legitimate can and will be debated until, in time, we acknowledge the equity issues covered by the smokescreen of technical questions and reach consensus on providing appropriate practice for all students.

The pedagogical issues regarding the most effective ways of preparing students for tests are, however, empirical and should be researched. Informal reports of experience with district-wide efforts to improve test scores provides conflicting evidence. Group achievement scores are clearly improving, especially in centers of low achievement. Higher scores stimulate undeniable short-run public relations gains that are difficult to challenge. There is much talk of superintendents whose contracts have been voted up or down, in large measure, based on how test scores looked. Certain children have progressed through a given grade, been placed in a special program, or just barely made it into the graduation line as a result of intensive "testing drills." The problem is that we do not yet know the answers to such fundamental questions as: What are the long run effects of extensive practice for basic competency tests? Does test-focused preparation contribute to or take away from broader-based achievement or from later success with more complex or more comprehensive tests? Furthermore, is there a difference in the effects of practice testing for students of different achievement levels, experience and backgrounds, language heritages, learning styles, or educational goals?

For a test score to be useful, it must represent a reliable sample of the student's actual achievement level. Clearly, the most representative test behavior will come from students who are comfortable with the test situation and confident in their ability to perform well. This comfort comes, however, only when students are well prepared for the tests they face, are motivated to perform their best, and have achieved a high level of skill in the tested subject and in test taking itself. We expect to teach the curricular content; integrating test taking skills with instructional content is a recent phenomenon. Nevertheless, test preparation that moves towards accomplishing these goals for all students is legitimate, expected, and long overdue.

## NOTES

1. The literature on test preparation, the effects of coaching, and test administration is extensive. I have listed some of the research I have found most useful in the accompanying reference list.
2. Most large city school districts and state offices distribute test administration instructions and test preparation materials annually. They can be obtained for review by contacting the state and local testing offices. Examples from Milwaukee, Los Angeles, and Broward County are cited in the References. In addition, the Florida State Department of Education has recently drafted a discussion document called, "The Standards for Comparable Testing Procedures," that outlines standards for collecting data on student achievement in schoolwide, districtwide, or statewide testing programs in Florida. I understand from informal conversations that similar standards design projects have been initiated in other states and locales. For more information about the Florida project, contact Dr. Barry Greenberg, Florida International University, Miami, Florida.
3. Ethical questions about test preparation at the district level were discussed in a 1985 American Educational Research Association symposium, Ethics, Politics, and Psychometrics in Detecting Irregularities in School Districts' Testing Programs, Ellen Pechman, Chair. Frary, Ligon, Perlman, and Stringfield & Hartman (referenced in the next pages discuss the issues in depth.

## References

- American Psychological Association. (1985). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.
- Anastasi, Anne. (1976). Psychological Testing (4th Ed.). New York City, New York: MacMillan.
- Bollenbacher, J. (1975). Standards for educational and psychological tests. Journal of Educational Measurement, 12, (1), 55-56.
- Broward County Public Schools. (1980, March). Improper and unethical testing practices. Fort Lauderdale, Florida: Evaluation and Testing Department, The School District of Broward County.
- Cole, N. (1982). The implications of coaching for ability testing. In Alexandra K. Wingdo & Wendell R. Gardner (Eds.) Ability testing: Uses, consequences, and controversies, pp. 398-414. Committee on Ability Testing Assembly of Behavioral and Social Sciences, National Research Council, Washington, D.C.: National Academy Press.
- Cronbach, L.J. (1970). Essentials of Psychological Testing (3rd Ed.). New York City, New York: Harper & ROW.
- Ebel, R.L. (1979). Essentials of Educational Measurement. Lexington, Massachusetts: D.C. Heath and Company.
- Educational Testing Service. (1983). ETS standards for quality and fairness. Princeton, New Jersey: Educational Testing Service.
- Frary, R.B. & Olson, G.H. (1985). Detection of coaching and answer copying on standardized tests. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.
- Gifford, C.S. (1981). Test-taking made easier: How to win the testing race. Danville, Illinois: Interstate Printers & Publishers.
- Gronlund, N.E. (1985). Measurement and Evaluation in Teaching. (5th Ed.) New York City, New York: Macmillan.
- Iverson, G. (1985, Winter). Appropriate test preparation: Do's and don'ts. Michigan Personnel and Guidance Association Journal, 16(10), 34-36.
- Ligon, G. (1985, April). Opportunity knocked out: Reducing cheating by teachers on student tests (Publication #84.36). Paper presented at the Annual Meeting of American Educational Research Association, Chicago, Illinois.
- Los Angeles Unified School District. (1982). Helping students do their best on standardized tests. Research and Evaluation Branch Bulletin (No.6). Los Angeles, California: Los Angeles Unified School District.
- Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. Educational Psychologist, 17(2), 67-91.

- Millman, J., Bishop, C.H., & Ebel, R.L. (1965). Analysis of testwiseness. Educational and Psychological Measurement, 25, 707-726.
- Milwaukee Public Schools. (1985, January) Guidelines for citywide standardized testing. Milwaukee, Wisconsin: Milwaukee Public Schools, Division of Planning and Long-Range Development, Department of Educational Research and Program Assessment.
- Nitko, A.J. (1983). Educational Tests and Measurement: An Introduction. New York City, New York: Harcourt, Brace, Jovanovich, Inc.
- Perlman, C.L. (1985, April). The Chicago test audit: A case for study. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.
- Sarnacki, R.E. (1979). An examination of testwiseness in the cognitive domain. Review of Educational Research, 49(2), 252-279.
- Stringfield, S.C. & Hartman, A. (1985, April). Irregularities on testing: Ethical, psychometric, and political issues. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois.
- Thorndike, R. & Hagen, E. (1977). Measurement and Evaluation in Psychology and Education. New York City, New York: John Wiley & Sons.

Discussion: You Really Oughtta Wanna  
Peter Wolmut

Multnomah OR Education Service District

Is it not fascinating to hear so many approaches to what one would think is such a straightforward topic? Covert performed a survey and found that a majority of respondents have no formal system for dealing with legitimate ways to prepare students for testing.

Ligon gives us Austin's 5-point program for supporting legitimate - and suppressing illegitimate -- ways of preparation, accompanied by references for pursuit of those points in more detail.

Matter, a long-time collaborator with Ligon in this arena, warns us that we need to share the body of knowledge about legitimate ways of preparation - if only for self-preservation.

Pechman also agrees that the legitimate ways have been well established. What isn't being recognized, she says, is the inequity between those who are continually "prepared" vs.. those who are lucky if they ever receive legitimate preparation.

Finally, problems and issues elicited from members of the audience also strongly indicate that there is a massive performance discrepancy between what is desired to be done and what is, in fact, getting done.

As long ago as 1970, Robert F. Mager and Peter Pipe proposed that one can detect such discrepancy through the phrase "You Really Oughtta Wanna". And that phrase kept coming silently to my mind so often in these discussions:

Teachers really oughtta wanna administer tests the correct way; teachers really oughtta wanna not teach the items on a test; teachers really oughtta wanna refrain from cheating (And, by the way, in one of his references, Ligon found that teachers tend to cheat on behalf of their students, not themselves!); administrators really oughtta wanna -make certain that legitimate ways of test preparation are promoted; test specialists really oughtta wanna, publish codes of ethics, etc.

Mager and Pipe suggested that part of the analysis of performance discrepancy is an examination to see if persons possess requisite skills. In 1984, I examined the testing requirements for teacher certification in the United States and found that 60% of the states do not require their students to take even a single course in test and measurements or research procedures. Part of the problem, then, is that most persons are not taught about legitimate ways of preparation in their pedagogically formative years.

At the same time, the problems cited here went further than what can be laid at the doorstep of the schools of education. What Mager and Pipe also pointed out a decade and a half ago was that performance discrepancy analysis requires one to examine what consequences occur as a result of the discrepant behavior -- positive, negative, or none whatsoever. And I believe this is where we must direct our efforts in our local districts when we leave here.

For if the purpose of all this testing is seen as pay raises, or the basis of the local newspaper's "Top Ten Schools", or that we have always tested, or because they should want to, the cited performance discrepancy will continue to sprout like the proverbial dandelion. On the other hand, if the purposes of testing are seen to provide positive consequence for the front-line teacher, and testing specialist, then I suggest the legitimate ways will indeed be followed.

### References

Ligon, Glynn "Opportunity Knocked Out: Reducing Cheating by Teachers on Student Tests", Paper presented at American Educational Research Association convention, Chicago, IL, 1985

Mager, Robert F. & Pipe, Peter Analyzing Performance Problems, (Belmont, CA: Fearon Publishers, 1970)

Wolmut, Peter "Issues & Problems in Testing in Big City Schools: Teacher Training in Measurement", Paper presented at the convention of National Council on Measurement in Education, New Orleans, LA, 1984

## Some Reflections On Legitimate Ways To Prepare Students For Testing

Brenda H. Loyd

University of Virginia

The comments and discussion presented by the participants in the 1986 NCME session on "Legitimate Ways to Prepare Students for Testing" reflect the current political emphasis on the need for accountability -- accountability of school systems, school buildings, and teachers -- and the assumption that the best way to assess performance in such an accountability-conscious context is to use the results of standardized tests as a measure of the performance of teachers, buildings, and school systems. The major topic under discussion here has been the problem of maintaining the integrity of the test in an atmosphere where teachers, principals, and superintendents have a large personal stake in the outcome of student achievement tests. It is natural, given this context, that there should be such intense interest in developing and enforcing guidelines for maintaining the security of such tests. But the security concerns which have been expressed here are merely one symptom of a much deeper problem which needs to be addressed. Much time, energy, and resources have been devoted to the development of security procedures, but

the cause of the security problem, and the effects of the policy of using student tests to evaluate educators, have not been given adequate attention.

My reflections on the current situation arise from the variety of perspectives from which I have experienced the development and use of standardized tests. My comments, then, reflect not only my particular area of interest as a university teacher of educational measurement, but also my experience with standardized tests as a graduate student at the University of Iowa, as a public school teacher, and as a parent.

During my years as a graduate student, I was fortunate to work on the Iowa Tests of Basic Skills. This experience gave me a first-hand understanding of the time, expertise and care that goes into producing quality standardized achievement tests. It also gave me insight into the impact of such instruments when they are used to enhance the education of our children.

The value of standardized tests has been above all to provide an objective assessment of student performance -\_ objective in terms of giving teachers and administrators an additional piece of information and additional point of view in judging the growth of students in the classroom. Teachers can use the results of such tests in a number of ways: to look for confirmation of growth or development; to look for

surprises -- the student who performs differently than expected (better or worse than performance in class would indicate); to resolve questions raised by such surprises or discrepancies -- i.e., to expect more of some students or to provide greater challenges for them, or to see if illness or stress reduced the performance of some students on the tests; to allow the teacher not to lose or overlook a student's ability or lack of ability because of personality characteristics such as aggressiveness or shyness.

Administrators can use these tests to evaluate the effectiveness of the curriculum. Plotting performance of students across the years allows principals or curriculum supervisors to see the relative strengths and weaknesses of programs for a system.

Standardized tests have worked extremely well for students, parents, teachers, and school systems (relative to other options). They have been effective in providing supportive evidence of teacher and school reports of performance, and in helping parents understand the strengths and weaknesses, and relative performance of their children. They have been found to be reliable, reasonably valid measures of student performance.

It seems to me that the main reason these standardized tests have been employed recently in accountability contexts is precisely because they have

been so effective in providing the information they were designed to provide -- because of their good reputation as objective, reliable, valid measures of student performance. However, the need for a session such as this one at NCME is evidence that such use of these tests has had unanticipated consequences. It is now clear that the testing process itself can be undermined by teachers who have such an intense personal stake in the performance of their students. No longer is the standardized test viewed as a tool that a teacher uses to obtain additional insights into students, or that school curriculum supervisors use to evaluate the effects of current curricula or changes in curriculum. Now the teacher can feel legitimately threatened by the use of such tests. Teachers are under pressure to see that their students are ready for the test, and this sometimes results in coaching students on test-taking techniques, putting students under pressure to perform at their best, excluding children from testing who may not perform well, and even "cheating on tests." In such a context, the results of the test actually become less reliable, valid, and objective -- that is, they begin to lose their effectiveness as accurate measures of student performance and lose their value as a source of information for the teacher.

The real problem is a simple one -- student achievement instruments are being used to measure outcomes that they were not designed to measure. If such tests are to remain effective indicators of student achievement, they cannot continue to be used to evaluate teacher performance or school system performance. The potential for confounding the results of such tests through the introduction of pressures and expectations which should not be a part of the testing experience is great, and the threat to the usefulness of such tests is genuine and lamentable.

This is not to say that student performance should not be a factor in evaluation of the instructional process at the classroom, building or system level. But for such purposes student performance should be measured separately. In fact, measures of student performance for use in the accountability contexts which are the focus of today's discussion could be obtained more efficiently and more effectively by procedures (such as matrix sampling as used by NAEP among others) which are specifically designed to obtain precisely the information which is needed for the evaluation of teachers, buildings, and systems. In the implementation of such procedures, it might well be appropriate to employ special precautions related to the security of tests and special regulations for administration which would reduce the potential for

"cheating." In a context where the main purpose of the testing is to provide information about the performance of groups (classrooms, etc.), the application of special guidelines to such measurement procedures could be taken into account in the construction and administration of the tests and in the interpretation of the results, and this could potentially be more cost effective than special security arrangements for the tests currently being administered. I would encourage the development and use of the approaches and instruments that most effectively provide the information needed to respond to the accountability question raised in our states.

But as a parent (and as a former classroom teacher), I do not want to lose the valuable tool that was created to aid the teacher in understanding the students. When my children take the standardized tests which are designed to provide important information about their progress, I want to know both their strengths and their weaknesses. I do not want the classroom at the time of testing to resemble an armed camp; I want it to be a comfortable, familiar environment for them. Furthermore, I do not want external pressures to be applied to our children to do well on the test in order to ensure a positive evaluation of the teacher, principal, or superintendent. It seems to me that if we allow this to happen or permit it to continue then we have defeated

ourselves. The reason for accountability is to ensure quality education, but if we remove the tools that the teacher needs we lessen the chances of improving the quality of the educational process and increase the chances of distorting the information that comes from standardized tests.

Across the country, substantial amounts of time, energy, and money are being spent to address the problem of maintaining the security of the testing procedure. But the reason for the problem is that the test is being used for purposes foreign to its intended use. My recommendation is that we use some of these resources to look for innovative new ways of obtaining information about performance above the student level (school, system, region, state). In the process of developing such procedures, standardized testing information could be used in a limited fashion to help clarify or validate the procedure. But in the long run, the two processes -- evaluation of student performance and assessment of educator or system effectiveness -- must be handled separately. Above all, we must preserve standardized testing at the student level as a valuable source of information about each student.

Symposium II  
Taming the Rasch Tiger:  
Using Item Response Theory in Practical Educational Measurement

This symposium was a response to requests from many test directors for a jargon-free discussion of item response theory in terms of its applications to their world. Since Georg Rasch first postulated his models in 1960, the subject has appeared to be quite esoteric, understandable only by a few privileged persons in the field.

Dean Forbes attempted to provide meaning to many of the terms and described practical methods of applying the Rasch model. Barbara Hunt explored the value of calibrated item banking in constructing tests closely related to the curriculum, while Clarence Mershon discussed the advantages of Rasch-oriented test development in refining a district's performance standards.

Gregory Thomas proposed that the multiple advantages of such methods have their own set of accompanying disadvantages. William Coffman discussed the preceding papers and shared a set of cautions

The Rasch Model as a Practical and Effective  
Procedure for Educational Measurement

Dean W. Forbes  
Portland OR Public Schools

Introduction

Education has always been faced with the problem of measuring achievement status and growth. In such measurement, testing people have always wrestled with the problem of matching each child with an appropriate test. It is all too apparent that problems occur when a child is faced with a test which is far removed from current performance level (either too difficult or easy). In either case there is frustration and, in the case of the overly difficult test, a great deal of pain. This problem of arriving at an optimum fit between a child's capability and test difficulty has been one that for years has plagued the testing field.

School Achievement Testing

There have been many attempts to improve the measurement of student performance both for individuals and groups. One of the great developments of the twentieth century was the establishment of tests for specific grade levels which could be normed on a supposedly national population. Theoretically this permitted comparing any individual child's performance with that of others at the same grade level using "national" norms. (It is not appropriate at this point to enter into any debate about the validity or utility of allegedly national norms.) The grade level test was a valuable improvement. It still fell far short of need, however, since many lower performing children found such a test completely beyond their ability.

In an effort to improve the situation there has been, in more recent years, a decided move toward out-of-level testing (where a low performing fifth grader might get a test originally designed and intended for fourth graders). This strategy created some problems in its own right since it then was difficult to interpret performance in light of norms either for the child's actual grade or the grade for which the test was originally intended. Various vertical equating efforts have attempted to provide a basis for interpreting out-of-level testing in meaningful terms.

If a school system is satisfied with the curriculum content of commercially published tests this might be an acceptable compromise to the testing dilemma. However, some residual problems still remain. Such tests, available from commercial publishers, are of necessity planned to handle all children at a given grade level (many test levels in fact span two grades). This means that the difficulty range of such a test is necessarily rather wide. There will be some very easy items to fit the lower performing student and there still have to be some very difficult items to fit the higher performing student; the average student for a given grade level can be nicely handled by the middle difficulty range of the test. Still, at any performance level there will be many items inappropriate to a given child. These non-functional items serve no useful purpose but do take additional testing time and increase the effort required on the part of the student.

If school districts embark on a program of local test development they will find that they have basically the same problems although they might not be as aware of them. One great improvement that is possible in district test development programs is planning and building tests to fit the specific local curriculum (rather than accept the compromises which a commercial developer must make). However, out-of-level testing will have the same problems as are encountered with commercially developed tests, and this may be further complicated by the fact that reference to any type of "national" norm is quite difficult.

Ideally, it should be possible to present each student with a test that is appropriate to that person's current performance. In other words, the test is of a difficulty level that can be handled with comfort. Furthermore, the test should have a range of difficulty narrow enough so that all (or at least most) of the items would be reasonable tasks for the particular child. If such a test system could be built to fit the local curriculum, the measurement millenium would approach.

In order to achieve this goal one would need a series of many tests (each of which would be short and built in terms of a restricted difficulty range), sequenced in order of difficulty, so that any child could at all times be fitted with a test appropriate to present performance capability. Although this need was recognized for years the test development and norming problems seemed insurmountable.

#### Recent Developments in Test Theory

In recent years new developments have emerged which address these problems. Unfortunately, they have brought with them new vocabulary and concepts which seem to alienate, rather than entice, the very user group which stands to gain the most.

"IRT," the "Rasch model," and "latent trait" are three terms which have recently risen to the status of "buzz words." They are frequently used in social situations whether or not they are really understood by the users. The somewhat better informed dilettante, like the naive wine taster, can simulate profundity by exercising a "smattering of ignorance." This might, for instance, involve reference to the "one parameter" versus "three parameter" controversy. This practice may easily intimidate those practical measurement people who lack extensive technical backgrounds and thus do a great disservice to one of the most dramatic and valuable developments in the recent history of test theory.

Rather than inform people about vital new testing tools, overly technical and jargonish explanations all too often becloud the relevant issues. This can create anxiety and lead to avoidance on the part of the very people who would most profit from the use-of these new tools.

There is nothing mystical or esoteric about these (and other closely related) terms or the concepts they represent. Let's find out what they are and how we can use them.

IRT simply means "Item Response Theory." It is a generic term which refers to any of a number of theoretical positions which relate to a mathematical model the probability that a person will correctly answer a test item. Such models occur in varying degrees of complexity. (A thorough exposition of this can be found in Lord and Novick, especially in chapters 17-20 which were authored by Allan Birnbaum (5)).

The Rasch Model is a mathematical model developed by George Rasch,, a Danish statistician. It is one of the models included under the general heading of I RT.

A Latent Trait is a stable and consistent characteristic underlying behavior which is, itself, unobservable. Knowledge of a person with respect to a latent trait provides a basis for predicting behavior. Examples of possible latent traits of interest to educators would be reading comprehension or arithmetic computation.

Now that we have defined some terms let's return to the issues and problems in educational measurement and see how we can cope with them. Let's find out how to apply Item Response Theory to their solution. Discussion will focus on the Rasch model itself.

#### Latent Traits and Their Measurement

Education has always tried to teach in terms of developmental continua. These have been recognized in the basic tool skills (such as reading, mathematics, and language usage) as well as in many more specialized instructional sequences.

It has always been held that certain stages are prerequisite to later stages. This has long formed a conceptual frame for education. As curriculum diversity emerges (after basic skill levels), each branch of instruction may suggest a possible continuum of its own while at the same time showing a divergence from other curriculum content areas. (Piaget spoke to the same issues in physical and cognitive development.) Thus, a conceptual map of education shows an hierarchical organization both within and between its various elements. Although teaching can be so organized, the learning continua underlying the curriculum have consistently been measured in rather haphazard fashion not only in the classroom, but by the traditional wide range standardized achievement test. In summary, this instructional model has been hampered by many measurement problems.

With acceptance of the premise that a measurement continuum should parallel a learning/ teaching continuum, it was recognized that if a learning continuum could be adequately measured by an underlying scale extending through its entire range, a student could enter and exit from the measurement continuum at points appropriate to that student's current development regardless of age or grade level. It would follow that scores for such continuum segments could meaningfully be compared either over time or between students. If this could be done, an extremely flexible educational testing system would evolve. The measurement of such an hypothetical testing model remained an unobtainable ideal for years.

In 1967 Benjamin Wright delivered a landmark paper at the ETS Invitational Conference on Testing Problems (10). This paper described a test development system based on one of the many measurement models developed by George Rasch (of Denmark) and his colleagues (8). This measurement model (popularly called The Rasch model), as described by Wright, seemed to promise a solution to many of the problems that we have been discussing since: 1.) "scores" were based on the curriculum difficulty of test items, not on the collective response of a comparison group; and 2) the difficulty of items was measured in terms of the probability that a person of given "ability" would get the item correct, rather than the proportion of persons in some reference group who made the correct response.

This model would permit selecting those items from a parent item bank with difficulty levels that matched the present ability of a student, it would further permit relating a "score" on that particular test to a measurement scale underlying the entire difficulty range represented in the total pool of items in the bank.

Thus, the specific model is based on one item parameter (difficulty) and one person parameter (ability). (Ability as used by Rasch and Wright, and in the present paper, merely means capability within whatever subject domain is being measured. It has absolutely no connotation of intelligence or any similar generic ability.) The model is probabilistic (i.e., it builds on the probability of "this person" getting "this item" right). It requires nothing else in order to arrive at a score describing performance but it does make some assumptions, and rather stringent assumptions too, about the nature of the items and of the educational element being measured.

The Rasch model assumes that all items will have equal (or nearly equal) discrimination and that all items involve minimal guessing. The model also assumes, with respect to the educational element being measured, an underlying latent trait of uni-dimensional nature. In other words what is measured is a continuum, not some composite involving content from two or more discrete or inter-twined continua.

The item assumptions must be met during the processes of item writing and bank development (and through test administration practices). The latent trait must be evaluated by fit of individual items to the model (probably after some preliminary intellectual analysis and soul searching).

Given a collection of items that fit the model a person may be tested with a selection of items appropriate in difficulty to that person's present capability and the raw score can be converted to a scaled score in terms of the scale describing the latent trait. Merely the number of items correct is a sufficient statistic for determining performance level.

This provides a tremendous amount of flexibility since a test could, in theory, be prepared for each individual student. It would also be possible to develop tests for particular performance levels and then administer those tests to all appropriate students. It would even be possible to develop wide range tests which could be used in similar fashion to those which have been a part of education for so many years. This latter procedure may at times be appropriate, although it will carry with it many of the problems discussed

earlier. It would be possible to generate short criterion-referenced tests to determine when a student reaches some particular performance level. In fact, so long as the items fit the model it seems to be a test development "Shmoo" (which, for those of you who do not remember the Al Capp cartoons of many years ago, was a "critter" that was all things and all people -- a truly universal specific to meet all needs.)

In practical test usage this means that any person can be measured in terms of present need, and that no two persons must take the same test. This would lend itself well to tailored testing either in the form of printed booklet or computer presentation. In group situations each student could get a "suit off the rack" (i.e., an appropriate and good fitting test from a pre-established series of test). (Over the past ten years the Portland, Oregon, Public School District has built an entire achievement testing program based on this model (4, 7).

There are a number of other latent trait models. A three parameter one has received much attention and study. It adds parameters based on item discrimination (slope of the difficulty function of the item) and the level of chance performance (guessing). The three parameter model is, of necessity, more complex than the one parameter model.

In general, the single parameter model offers some very real advantages. In the process of developing banks of appropriate test items and validating them for latent trait measurement (a process called calibration) the three parameter model is complicated by the fact that item calibration requires larger samples of respondents than does the one parameter model (which can satisfactorily calibrate an item with 200 or fewer responses (7, No. 20)). Several computer programs are readily available for the calibration of items for the one parameter model and some of these make modest demands for computing power (16). The one parameter model has proved to be very successful in a wide variety of practical educational situations. It has been used by the public schools in Portland Oregon, other member districts of the Northwest Evaluation Association, school districts in California (particularly San Jose), in England, and in Scotland (3,6,7,9). The basic procedures involved in the calibration of items using the single parameter Rasch model are described in Wright's Invitational Conference Paper (10), and Wright and Panchapakesan (15). The basic procedures for building and using tests from item banks are described in two research memoranda from the Statistical Laboratory of the Department of Education, University of Chicago, by Wright and Douglas (13, 14). A more complete presentation is given in Wright and Stone (16).

#### Item Generation and Item Bank Building

An essential ingredient in employing the Rasch model (or any other IRT model) in test construction is a pool of relevant test items which have been demonstrated to fit the model. This means that the items have been calibrated (a process which gives all of the traditional item analysis information plus information verifying fit of the item to the Rasch Model). Computationally the calibration process is laborious but straight forward. Current micro-computers can handle calibration and appropriate programs are readily available. Among those available are BICAL,, which deals with the one

parameter model, and LOGIST, dealing with the three parameter model. (In Best Test Design Wright and Stone outline a procedure for hand calibration of items which is actually practical for small situations (16).)

Assuming the availability of skilled item writers, it is possible for a school district to generate its own items. If the district has access to an appropriate computer, it can also calibrate them. If a school district is not able to do this by itself, it is entirely possible to join forces with other interested, school districts to pool resources. The Northwest Evaluation Association is an example of such a consortium. It involves independent school districts, primarily from the States of Oregon and Washington. The association also, includes representation from California and various other states.

If one does not wish to embark on an item bank development project (with its necessary item calibration) there is a steadily increasing number of Rasch calibrated item banks in existence (some of which are available for purchase (12)). One well established set of item banks (in reading, mathematics, and language skills) is available for purchase from the Northwest Evaluation Association.

An item bank can occur in more than one form. Each of these has certain implications for use. In the simplest form it can even be a file of item cards without an index (to preserve the user's sanity any bank larger than 50 or 60 items needs an index, and even those that are smaller should have one). Large files of item cards can be managed manually with adequate indexing. Indexing and index referral are greatly facilitated by computerization; this permits a great saving in clerical time by letting the computer identify those items meeting, the test constructor's specifications. In this way, all non-relevant items can be ignored. (A strictly clerical application would involve scrutinizing each item card to determine whether or not it met the requirements specified by the test author. It doesn't take much imagination to see that this could become a very time consuming and laborious process.)

A higher stage of item bank development puts the items themselves in computer storage (either on hard or floppy disks). For many years practicality limited computer item storage to textual or numerical material with graphics being stored, by necessity, in a supplementary card file. More recent developments, however, permit computer storage of graphics and the introduction of laser beam technology promises even greater sophistication in the near future.

#### Item Bank Usage for Test Assembly

Once a user has access to an item bank, a wide variety of tests can be constructed to meet many measurement needs. The process of planning and assembling a test will follow basically the same steps with which you are all familiar. Goals to be measured must be identified, the number of items per goal must be decided, and the specific items to fit this blueprint must be selected. (Remember, we are dealing with a bank of items already written and calibrated. If the contents of the item bank cover the goals to be tested it will be necessary to generate few, if any additional items.) Now, however, you can add something new to the process. You decide the level of difficulty which you want the test to measure, and the range of difficulty it should

span. At this point testing is being done "on purpose" instead of "by accident." It is even possible to build a series of tests to the same blueprint but with planned differences in difficulty level. This permits accommodating individual differences in a way previously not possible while still preserving the ability to convert raw scores from different tests to a common scale. (This is the city-wide achievement testing model that has been developed and is in use in the schools of Portland, Oregon.)

Tests, regardless of purpose (i.e., unit or semester examinations, criterion referenced achievement tests, or survey achievement tests), can be assembled using a variety of procedures.

They can be put together by purely clerical means. This would involve selecting appropriate item cards which would then be turned over to a typist for preparation of final print copy.

Item cards can be selected, superimposed in page format and placed in an acetate sheath for photocopying of "print copy" pages (this substantially reduces the possibility of clerical error in item reproduction).

Once items are computer stored it is possible for them to be selected and displayed on the computer screen, or printed out by purely electronic means. This permits a number of possibilities. All items can be listed and displayed for final editorial checks and selection. With adequate computer programming it permits actual computer printing of "print copy" pages. It makes possible (again with adequate computer programming) computer assisted testing, whereby the computer presents the student with one item at a time on the video screen and the student uses the computer keyboard as a response mechanism.

This last procedure offers several alternative procedures with differing degrees of sophistication: 1) computer presentation of one pre-constructed test; 2) computer presentation of one test selected from a group of pre-established tests; 3) a tailored test where the computer "jumps" a person to an appropriate item difficulty range and then gives a pre-selected sequence of items; or 4) a computerized adaptive test where the computer initially establishes an estimated performance level and then individually selects each succeeding item based on the correctness or incorrectness of the previous response. In this case testing would involve as few items as are necessary to document the performance level at a predetermined level of measurement error. At that time testing would be ended.

The Scottish Ministry of Education has subsidized the development of some very promising item banks and test construction procedures which have been built around microcomputers and related equipment (2, 3). These projects are programmed so that the computer selects items to meet blueprint specifications set up by the test author. In selection of items the computer program has available various default procedures that can be used in the event that item substitutions must be made due to incompleteness of the basic item files. The computer then formats and prints the actual test pages (including graphics) so that all that is left is the printing. At that point the teacher (or other user) has a test available which fits the-measurement needs of the specific situation and which is composed of high quality items. And all at minimal expenditure of effort.

Computerized item bank and test development facilities of this sort can be developed so that they are strictly "stand alone" operations or they can be developed so that a remote station can interact with a larger central processor (and item banks) by means of telephone circuitry.

#### Summary

For many years personalization of achievement testing has been impossible in all but the simplest forms. Recently, item response theory has emerged as a valuable tool which brings far greater flexibility to the process than had previously been possible. The single parameter Rasch Model has proven particularly useful in a number of instances in which it has been used. These situations range from individual school districts, through consortia of school districts, to governmental agencies.

IRT makes possible the measurement of performance in relation to latent traits in a way which is especially appropriate to basic skills achievement testing in the public schools.

Recent developments in micro-computer technology and the development of computer programs, make it practical for the local school district to develop and calibrate files of items to meet its own testing needs. Calibrated item files are becoming more available for purchase by those potential users who would sooner not develop their own.

The transition from item file to finished test is well within the scope of the local user. Available procedures range from the intensively clerical to the highly computerized and are well documented in readily available publications.

### References

1. Bagnall, G.M., Milne, P., and Pollitt, A.B., Primary Mathematics Item Bank (Explanatory Manual). Scottish Education Department, Godfrey Thompson Unit, Department of Education, University of Edinburgh (1982).
2. Bagnall, G.M., Milne, P., and Pollitt, A.B., Primary Mathematics Item Bank (Summary Report). Scottish Education Department, Godfrey Thompson Unit, Department of Education, University of Edinburgh (1984).
3. Godfrey Thompson Unit, Item Banking Facility (Introduction). Godfrey Thompson Unit, Department of Education, University of Edinburgh (Feb. 1984).
4. Kingsbury, G. Gage, A Comparison of Item Response Theory Procedure for Assessing Response Dimensionality. Paper presented at the annual meeting of NCME, Chicago, Ill., 1985.
5. Lord, Frederic M., and Novick, Melvin R., Statistical Theory of Mental Test Scores. Addison-Wesley Publishing Company, 1968
6. Martois, John S., Measurement Issues Related to the Use of Item Response Theory Based Life Skills Reading Tests. (Personal communication from Dr. John Davis, San Jose, California, Unified School District.)
7. Occasional Papers, Portland (Oregon) Public Schools Department of Research and Evaluation (various authors, topics, and dates).
8. Rasch, George, An Individualistic Approach to Item Analysis. Readings in Mathematical and Social Science. Lazarsfeld and Henry (eds.), Chicago: SRA Inc., 1966.
9. Rentz, Robert R., and Bashaw, W.L., The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Educational Measurement, V14, No. 2 (Summer 1977)
10. Wright, Benjamin D., Sample Free Test Calibration and Person Measurement. In invitational Conference on Testing Problems, Educational Testing Service, 1967.
11. Wright, Benjamin D., Solving Problems with the Rasch Model. Journal of Educational Measurement, V14, No. 2, (Summer 1977).
12. Wright, Benjamin D. and Bell, Susan R., Item Banks: What, Why, and How. Journal of Educational Measurement, V21, No. 4 (Winter 1984).
13. Wright, Benjamin D., and Douglas, Graham A., Best Test Design and Self Tailored Testing. Research Memorandum No. 19, (June 1975), Statistical Laboratory, Department of Education, The University of Chicago.
14. Wright, Benjamin D., and Douglas, Graham A., Better Procedures for Sample Free Item Analysis. Research Memorandum No. 20, Statistical Laboratory, Department of Education, The University of Chicago.
15. Wright, Benjamin D., and Panchepakesan, Nargis, A Procedure for Sample Free Item Analysis. Educational and Psychological Measurement, VZ9, No. 1. (Spring 1969).
16. Wright, Benjamin D., and Stone, Mark H., Best Test Design. MESA Press, Chicago, IL. (1979)

# Implications of Rasch Calibrated Item Banks for Measurement of the Locally Planned Curriculum

Barbara Hunt

Hopi Tribe Educational System

## A STATE LEVEL VIEW OF SCHOOL DISTRICTS

A view of local school districts from the State Department of Education level reveals districts that vary in many dimensions. These include cultural and religious composition; land wealth; geographic configurations; political persuasion; and citizen awareness. These differences are reflected in the politically powerful issue of maintaining local control of schools. Woe unto the state school superintendent who does not support local control of schools.

The size of a district is a crucial factor which few State Department of Education employees can overlook. Most western states are composed of a few large and medium size districts and a myriad of small, primarily rural, districts. Some of the large urban districts virtually control state level decisions by the legislature and the Department itself. One characteristic of the large districts that may exert so much control is population. This means votes, and is a crucial factor in school funding formulas. These result in large amounts of tax money and are invariably accompanied by regulations and legions of financial watchdogs. Large concentrations of population demand specialized services and programs staffed by specialists and professionals in addition to the normal configurations of teachers, principals, and Superintendent.

The hundreds of small districts are often characterized by a "fierce determination to maintain local control" and the will to "do it themselves.

Lack of population density often means fewer 'votes and less political influence. This, in turn, can lead to less money generated in funding formulas. Specialized services are still needed but small numbers of clients don't justify the numbers of service specialists that are necessary. As a result multi-talented or generalist personnel are better suited to the small school situation. In many states rural school administrators are organized into associations which help them handle the excess costs generated by small numbers requiring specialized services. Personnel in small districts often wear "many hats," and are generalists (as compared to the specialists found in large districts). The needs, however, remain the same.

Rural educators tend to be more in touch with members of the community because they are not insulated by the layers of secretaries, specialists, and other personnel who in large districts absorb and deflect the ire or gratitude ,of citizens. Rural educators can't hide and they demand answers they can implement with a minimum of time taken from the primary function of schools-which is the education of students.

## MEASURING THE MAIN PRODUCT OF SCHOOL DISTRICTS: STUDENT ACHIEVEMENT

Excellence in schools and education has become a political issue which has led to the selection of a teachernaut who tried to orbit the earth, and has stimulated nearly 50 state legislatures, which are involved in mandating changes in education in hopes of implementing educational excellence. No longer are educators the sole decision makers in education, local otherwise. The public, through its governing bodies and elected politicians, is asking hard questions. These are questions that can no longer be answered by educators on the basis of the same educational theory which governed decision making in the past.

The public and their spokesmen are demanding hard facts, data, measurement, and proof that the students are deriving educational benefits commensurate with the investment of public monies in school support. Title I ESEA and Chapter I ECIA first felt the demand for measurement and accountability. They responded by designing a system to measure the achievement growth of disadvantaged students which resulted from the congressional investment of approximately three billion dollars annually since 1965. Actual evaluation began in the middle 1970's and the positive results were such that they convinced congress at least once in the decade of the 70's -to renew the investment.

State Departments of Education took the brunt of carrying out the Congressional mandates. These included the use of nationally normed tests to measure achievement in local school districts where disadvantaged students had ,3-been served with Title I or Chapter I funds. State departments also carried .--out state legislature mandates for proficiency testing at certain grade levels.

Steadily, for a decade, the politically sensitive ears of State Departments of Education heard cries of alarm from local districts who pointed out that nationally normed and published tests did not really measure what districts had decided to teach. Nor did local educators feel they should give up local control of curriculum to teach. the content or language that publishers used in their test items which were intended to measure student achievement.

State Departments strongly suggested that districts should test what they taught, but had to dismiss teacher made tests as a viable method because of the many problems associated with norms (national or local) and the aggregation of data across many populations within the United States.

Today, there is an alternative which may resolve the controversy between nationally published and locally developed tests. The alternative lies in the use of Rasch calibrated item banks to assess, measure, and evaluate student achievement.

And that alternative is just in time. State legislators are insistent in their demands that educators at all levels assess student populations, measure ,achievement, and report student gains (either positive or negative).

Legislators, faced with economic shortfalls which affect state budgets and local school support, are, again asking if they are getting their money's worth in terms of student achievement. Predictions are that, although the questions now want answers in terms of basic skills achievement, down the road there may be demands for accountability as to the role education plays in solving societal ills such as unemployment., economic, growth, cultural bias, juvenile - justice, and a host of other issues

### THE CASE FOR LOCAL SCHOOL DISTRICTS AND ITEM BANKS

If, as stated earlier, the greatest percentage of school districts are experiencing crises in specialist/generalist personnel issues, as well as in population base/financial structures, who in a local school district is going to have the time, the awareness, and the commitment to become knowledgeable as to the nature and use of an item bank? And for what purpose?

Probably few if any school district personnel will be "commissioned" to find an item bank whether it is to be used by the district for task assessment evaluation, or as a teacher resource for instructional evaluation. However: just as I draw a hopeless conclusion, inevitably, a small cadre of educators steps forth. These seekers are sent by forward looking superintendents who have salvaged a small margin in the budget to finance such a search or by educators whose vision and need for better ways of doing things have galvanized them into self-actualizing searches for those better ways.

Rasch techniques and use of item banks are a "better way." They lend themselves to involvement of educators in evaluation procedures and use of the results that is unprecedented in the field of educational evaluation. Rasch tools can implement the flexibility of item banks. Such banks can adapt items to local language and conditions. The processes of scoring and score management are readily adaptable to the computer technology available in virtually any school district (even if it is limited to microcomputers).

Flexibility of item banks, equating scores to local and/or national norms and the potential for processing data on existing equipment make it possible to implement Rasch procedures with moderate financial outlay and maximum potential for staff involvement and use.

Teachers can select items which reflect their instruction, schools can select items which align curriculum and guide goal setting. The community can match student achievement and community goals, and school administrators can monitor instruction as it occurs. By providing relevant, useful, measurement schools can identify what they are doing and document how well they are doing it.

### THE CASE FOR RASCH CALIBRATED ITEM BANK USE IN ONE SCHOOL DISTRICT

The proceeding section is essentially an abstract of the salient point-s-and potential benefits of use of Rasch calibrated item banks. These points and theoretical benefits are merely academic until a clarification of the need

and the "fit" of the system to a specific district makes the case in such a fashion that the possibility becomes an indisputable fact.

The exemplar school district is currently a Headstart to grade 8 elementary school system on the Hopi Reservation in the Northeast corner of Arizona. During the school year 1986-87 the system will expand through grade 12 with the opening of a new high school in the fall of 1986. All but one of the elementary schools and the high school are administered by the Bureau of Indian Affairs and must comply with federally mandated regulations. The other elementary school is contracted by the Tribe and is run by a community elected school board.

The schools are each much like unified local districts under one superintendent with local village school or governing boards in addition to a reservation wide school board. The governing boards are fiercely proud of their autonomy and the villages view the village school as an integral part of the community.

With the advent of a new Jr.-Sr. High School on the reservation, high school-age students, grades 9-12, will no longer be sent to boarding schools. They will instead be bused from villages each morning and afternoon to the reservation high school. Hopi High grades 7-8 will be bussed from current Village grade schools to attend the Jr.-Sr. High School. Sending grade 7 and 8 students to the high school will mean a loss of enrollment in village schools and the possible closure of some schools due to lack of students. This could create a situation where a school would not be cost effective to operate--a major concern of government funding sources and now a major concern villagers.

The rules and regulations which provide the reality within which these schools must operate are as follows:  
FEDERAL REGISTER 36.30 Standard -X-Grading requirements - Rules and Regulations.

- (a) Each school shall implement a uniform grading system which assesses a student's mastery of the prescribed objective of the courses of study undertaken. The mastery of prescribed course objectives shall be the primary measure of academic attainment for reporting student grades on report cards.
- (b) The information derived from student instructional evaluations shall be shared with the student and with the parents and shall be used to give teachers and students direction for subsequent learning activities.
- (c) Parent/teacher and parent/teacher/student conferences focused on the student's instructional progress and development shall be held, where feasible and practical, to provide an additional means of communication between home and school. Residential schools may meet this standard by documenting the communication of student grades on report cards to parents.

36.3] Standard XI -, Student promotion requirements.

(a) Each school level or equivalent shall have a minimum criterion for student promotion based primarily on measurable mastery of the instructional objectives.

(b) Criterion-referenced tests that evaluate student skills shall be utilized for measuring the mastery of instructional objectives. The evaluation results shall form the basis for the promotion of each student.

## II. Village School Boards and Citizens:

1. Help determine the curriculum of the village school.

2. Want to measure progress of student learning on village goals for education and on national norms.

The advantages of the Rasch Item Bank System that follow are situation specific and atypical when compared to most school districts and their populations. A great difference is the feeling of the community members about the village school and their district involvement with all facets of the school. Village boards select teachers, advise teachers decide discipline policy, and assume responsibility to assist in discipline. The community offers, to most students, the nuclear family, the extended family, and the clan system to guide, discipline and care for the children. The culture's ceremonies, and different languages of home and school which exist for many students all serve to contribute to an atypical school setting.

### ADVANTAGES OF RASCH ITEM BANK SYSTEM

1. Village schools may develop a curriculum which the community can support and then can construct tests to measure student achievement as they progress through that curriculum.

A. They can build or purchase Item Banks from which to choose items relating to the goals they have taught.

B. To implement criterion referenced testing items may be clustered around school goals.

C. Calibrated items provide both individual and group scores.

D. Cross linking with national norms may be developed for a broader comparison.

2. Teachers may select items they feel are "functional level" to measure achievement with less chance of students "topping" or "bottoming" out of the measurement range of the items. This approach substantially improves measurement of individual student achievement.

3. Teachers and community can write their own items and "calibrate" them locally and/or district wide.

A. Local language (English dialect) can be used.

B. It is conceivable that parts of the cultural curriculum can be measured eventually.

Microcomputers and appropriate software are available to score basic skills tests.

If teachers find a "slow group" of students they can identify test items they have taught and refer items not taught to decision makers who can "adjust" curriculum.

Teachers will know then what previous teachers had not had time to teach. Goals, expectations, and objectives at grade levels can be adjusted without the guesswork inherent in conventional systems. This improves accountability of teachers and identifies grade levels that may need extra help or concentration of resources.

Items using "local language," thinking, relationships, etc. can be calibrated, compared with traditional items and used as a basis for deciding instructional strategies and language that is most effective.

In the Criterion Referenced approach selection of items may be made relevant to community goals and they can be measured in addition to other school goals.

Conceivably, items that measure science, social studies, and other instructional fields can be developed.

#### Disadvantages

1. Measurement of specific goals is limited to the comprehensiveness of available item banks.
2. Teacher judgment of a student's functional level is subject to error until the teacher has previous test scores to use in predicting present performance.
3. Calibrated item banks occur only in certain areas of the curriculum (mostly in the basic skills). Item bank development needs to proceed in other fields such as science and social studies.

Many problems are encountered by school districts regardless of size. In cases where population density and resulting funding fall below the level which permits an adequate specialized staff to provide curriculum and measurement services imaginative staff utilization must be used to fill the gaps. One procedure that is useful is to recruit generalists who can serve more than one function.

State and federal involvement in education provides certain financial resources, but also makes demands for accountability. This creates conflicts between local program development and evaluation, and the requirements of state and federal regulations. This particularly complicates the measurement of student progress.

Recent developments in measurement using the procedures developed by George Rasch provide great advantages in the flexible and comprehensive measurement of student achievement. These same procedures provide a means to equate local measurement to the results of nationally distributed commercially published tests.

Since the use of calibrated item banks offers far more flexibility for the measurement of student achievement than any other system available at the present time, it is especially appropriate to a school system as deeply imbedded in the life and culture of the communities as is the Hopi system. This program demands choices and flexibility that teacher and community involvement can influence and adapt to meet their needs.

In the Hopi tribe, village committees and boards have a strong role in what the school does and how the students act regarding education. These voices have echoed for over 1,000 years across the Mesa tops regarding education of the young people and these same strong beliefs are voiced today. Current Effective Schools literature heaps accolades on parent involvement as a cornerstone of student achievement. The Hopi tribe has parent involvement and now it must develop a responsive and flexible measurement system that will meet its educational needs.

## HOW PARKROSE IMPROVED ITS TESTING PROGRAM USING THE RASCH MODEL

Clarence Mershon Parkrose OR School District

My presentation will cover the development and refinement of the test Parkrose School. District since 1973. In particular, I will present data showing which we are using item banking and the associated curriculum scaled to improve our programs.

Parkrose School District's goal-based curriculum and measurement program developed as a result of a change in Oregon State Department of Education standards relating to education requirements. In 1969, the Department surveyed Oregon's needs by sampling opinion of students and dropouts, educators and the public. The survey revealed concern that the high school diploma had lost its credibility and that not all students were receiving training and instruction necessary to cope with the demands of society. The first draft of the resulting new graduation requirements, based upon student outcomes, was distributed in the fall of 1971. After review by interested individuals and groups, the new standards were adopted in 1972, and were to affect the graduating class of 1978.

In Parkrose, a steering committee was formed in 1972, and developed plans and an implementation timeline. In the fall of 1973, a Project Director was appointed to manage the effort. Planning involved administrators, teachers K-12, students and parents. Developmental teams were formed and were given responsibility for the curriculum planning at the program level. As the general framework for meeting State requirements emerged-committees were formed in program areas (such as mathematics, reading, science, etc.).

From the program's inception, District leadership perceived the State mandate as an opportunity to re-define curriculum in certain critical areas and to develop a program to monitor student performance in those areas. Coincidentally with goal-based curriculum definition, criterion referenced tests were developed to assess student performance in reading, language, mathematics and science. As these developments progressed, issues surfaced that required attention and resolution:

- (1) District Curriculum vs. "Teacher" Curriculums: This issue relates to the fact that a district adopted curriculum does not insure its use in the classroom.
- (2) Items vs. Curriculum: This issue concerns the alignment of test items to the defined curriculum. Do the test items used measure performance vis a vis; a particular goal?
- (3) Number of Items for Goal Coverage: How many test items are needed to cover reliable information about performance in relationship to a particular goal?
- (4) Timeliness in Reporting Results: In Parkrose, this seemed to be a problem. With tests, scoring and reporting took two weeks or more. We decided to reduce the time for scoring/reporting test results.

(5) Test Security-. What steps can be taken to eliminate teaching to a test?

(6) Standards: What is an appropriate way to establish criteria for a "Passing score? How does one ensure that standards are not too easy or too difficult? How can one be certain that parallel test forms are of a near equal difficulty level?

Each of these problems was addressed as development proceeded. Their resolution, I believe, came about because District. personnel were pragmatics, innovative in many respects and responsive to concerns of our constituency - teachers, students and parents.

It quickly became apparent that the District needed to do more than program definition. A District "curriculum" was extant, but teachers obviously did not feel bound to follow it. It established broad, general goals with neither specificity in outcomes nor direct measurement of achievement with respect to that curriculum. Accountability really did not exist except, at a most general level. Fortunately, the State mandate provided the impetus needed to help overcome reluctance to make the system more accountable. Starting in 1973, the curriculum was defined in terms of goal statements, a practical, intermediate level between the prior broad, generalized curriculum statements and what we perceived to be a minutiae of very specific behavioral objectives defined by some districts. In order to focus attention upon these goals, tests were developed, where appropriate, to measure achievement with respect to defined learner outcomes.

The next problem involved Ending test items which were referenced to the more specific goals established. At that time, a coarse goal collection published by a local Educational service District (Multnomah Educational Service District - M.E.S.D.) was available, but item collections, referenced to a system of goals did not exist in useable form. Therefore, District development, teams were paid to write items for specific goal & A special project in 1974, under the auspices of the Oregon Association for Supervision and Curriculum Development, provided an opportunity for staff members to be trained in writing items. High School. classes graduating under the PC= standards (d of 1976& 1977) were the subjects for Md. testing these items. Decisions were then required concerning the number of items per goal, the number of goals to be tested, the criteria for "Passing" and similar issues.

Originally, many teachers believed that the standard should be 100%, but results from the field quickly demonstrated the unreasonableness of such a standard. Also, the field tests helped us in eliminating redundancies, (where one skill Js dependent upon another). The District established the performance standard at 80% for most skill areas, which implied. some multiple of five test items for each goal area. Since reliability increases w3th the number of items, we established a minimum test length of twenty items, whether measuring one or more goals. The information available, together with = own follow-up studies, convinced us that this number (20 items) is necessary. The availability of the Rasch scaling system provides us with another possible standard for "passing" (though we do not use it at this point).

One factor negatively affecting the use of test results in Parkrose prior to this project was the time span between testing students and receiving test results from M.E.S.D. Generally this was two weeks or more. Also, the ability of teachers to analyze, synthesize and/or use the results for evaluative purposes was limited. Many questioned the content validity of standardized tests as well. The latter argument, of course, did not apply to our tests. To deal with the timeliness issue, a decision was made to report test results back to the teacher and student within twenty-four hours. This commitment enabled teachers to use the results immediately. In retrospect, this policy of quickly providing test data to teachers certainly helped convince them of the usefulness of test results. Teacher reports concerning the impact upon student motivation were universally positive. Also, teachers were most appreciative of our practice of maintaining all performance records, rather than requiring them to do so.

Another issue that arose very soon was the problem of test security. The District made a decision to have a testing coordinator administer the tests. Neither teachers nor students were to have access to the test prior to the time of administration. However, during the initial year the test coordinator reported previous tests posted in some rooms, and instances were reported of teachers using prior tests as a teaching tool. As a consequence, the Director started the development of an item collection with specific reference to District goal areas. Concomitantly, a system for utilizing a bank to publish print-ready copy was created. Shortly thereafter, the Northwest Evaluation Association (N.W.E.A.), a consortium of Pacific Northwest School Districts, came into being to provide forum for test development, field testing and test research. As a participant in that effort, Parkrose contributed test items which were field tested and calibrated using responses from students throughout the states of California, Michigan, Oregon and Washington, and which were added to the N.W.E.A. item collections. Because of the possibility of being able to construct parallel test forms at a given level of difficulty, the School District was eager to cooperate fully in this developmental work. Use of the collections and our ability to calibrate new items makes it possible to develop District tests covering the desired content at the desired level of difficulty.

This latter issue has provided the foci for our efforts the past few years. Over the past twelve years, our testing program has changed from a mostly "hand" to a mostly "automated" system. The availability of microprocessor software which provides a student data base, and which gives us the capacity of applying Rasch and Wier test analyses techniques enhances our work in this area. Mostly our goal-based testing program focused on graduation competence. As this program started to impact student development scores, the District decided to expand the program to include all levels. Changes in State standards in 1976 gave impetus to this effort, since these new standards required Oregon to monitor student performance in basic skill areas (reading, language and mathematics), K-12.

Our student data base provides the flexibility to test one time during the year and during the fall. The T.LP. student database enables the District to maintain records and monitor student growth from year to year, grade by grade. We use the database to calculate graphic profiles of achievement growth for individual students and groups of students from year to year (See Appendices A and B). These data provide the foundation for the evaluation of school and District instructional programs. Results from goal-based and competence tests are similarly compiled and reported as needed.

The database also provides an effective method for calibrating new test questions needed to assess District goals. We use the Fixed Parameter method to match the T.LP. results to District developed tests and calibrate the items to the N.W.E.A. scales in reading, mathematics and language usage.

This past spring (1985) student responses on District developed tests and the T.LP data were used to calibrate many of our District tests. Appendix C and D show an example of the fixed parameter calibration and scaling for the third and fifth grade reading tests respectively, Similar analyses were performed on the third grade mathematics test, the sixth grade mathematics test and a mathematics competency examination.

District personnel have been concerned that some of our tests were either too easy or too difficult. The Director of Program Evaluation predicted that both third grade tests (reading and mathematics) were too easy and the fifth grade reading test too difficult. Table 1 contains the comparisons of the average RIT level of students as measured by the T.LP. tests and the District's 80% standard established for each of these. As shown by the data, these three predictions were confirmed by results from the Fixed Parameter analysis.

Table 1

District Goal-Based Test	Pasch Level-80% Standard	District Average Rasch level
Third grade reading	190	194.9
Third grade mathematics	187	193.8
Fifth grade reading	232	208.2
Sixth grade mathematics	222	223.1
Test 1, Competency (Graduation) Examination	229	230.6

\*This test is given only to those 6th grade students who meet all criteria on the 6th grade mathematics test.

This research has made it possible to modify these tests to ensure that the 80% performance standard accurately reflects the achievement level desired. In the future, items from these tests can be used together with those in the item banks to construct parallel forms, expand the item collections and improve all tests.

Parkrose faces a dilemma similar to that experienced by other districts trying to improve its assessment program. Teachers are rightly concerned that tests measure performance vis a vis the curriculum. On the other hand, a district cannot long operate in isolation from what is occurring in other districts, hence the need for standardized measurement. Developing a system to accomplish the former without jeopardizing the latter is essentially what Parkrose is about. We also believe that student achievement scores will improve as the measurement and reporting systems improve, as the District is able to establish more appropriate performance standards.

## APPENDICES

A 1 to A 3

B 1 to B 6

C 1

D 1



## INFORMATION NEEDS WITHIN A MULTI-DISTRICT ENVIRONMENT

Gregory P. Thomas Washington County OR Education Service District

The Northwest corner of the United States is populated by a large group of professionals who would have one believe that creating a curriculum matched, difficulty indexed, individually oriented test was relatively easy. They point out the decade-plus history of item bank development, and the availability of inexpensive computer hardware as well as software.

An equally vocal group of professionals will point out the need to meet state mandated testing requirements, amortize investments in nationally normed test instruments, stay compatible with long data histories of student performance, and the importance of statistics and reported results which can be understood and explained to diverse audiences. Add to this mix the "new" thrust on school effectiveness and "profiling" and the seemingly simple issue of data and information becomes very complex.

I intend to argue in this paper that the simplistic approaches which many people urge will not produce information of utility in the management of the complex environment known as a school district. I will also explore the single data point notions prevalent today and suggest an alternative or two which might be used as one attempts to develop usable data for management (teachers also manage under my notion of management, not Just administrators).

### ITEM RESPONSE THEORY (IRT) - RASCH MODEL - LATENT TRAIT

Although each of these terms is definable and much writing has occurred (Birnbaum, 1968), for convenience I will deal with them as a single conceptual topic. The basic test development procedure used in these approaches begins with the development of test items by a team of "trained individuals." These individuals write items. The items are then put into a test format and data are collected from an examinee sample of convenience. One of a number of statistical treatments is then applied to identify items which do not meet the constraints of the particular statistical model being employed. Items are then either discarded or re-written.

This process is straightforward and familiar to all who have developed tests. The difference is that in using item response theory (IRT) the items are assigned a "difficulty index" based on their relationship to the other items in that testing session. The index is to serve from this point on as a principal key to item identification and use in future test development.

It is worth taking an abbreviated look at some of the techniques which are used in this process as well as the assumptions which must be adopted. A brief look at the Rasch model, one widely used in our area, will illustrate many of the assumptions. This particular model statistically contrasts curriculum difficulty of test items with the probability of an individual of certain ability answering the item correctly. "Good" items will show high discrimination if they are to be retained. In short, if a person is of "high ability", the item should be correctly

answered and, at the same time, incorrectly answered by a person of low ability. Whereas the Rasch model represents a single parameter model, much has also occurred as a result of the introduction of multiple parameter models (Lord and Novick, 1968). Multiple parameter models are akin to the exercise of adding variables to a multiple regression so that the error term is further and further refined in an attempt to enhance accuracy of the measure.

The samples of individuals used to generate item data are most typically "samples of convenience." Classrooms of students and, in some cases, school districts have been utilized as test sites for new items. The size of needed samples is accepted as 200 based on work done by the Portland Public Schools Evaluation Department (occasional paper, P.P.S.). Because of the pervasive belief that item difficulty does not vary across samples of differing ability, more traditional sampling approaches are not typically used.

However, holding the item pool constant but varying the examinees produces a somewhat different result. Reviewing student performance data across a two year span revealed that item difficulty will vary between administrations of a constant set of indexed items (occasional paper, MESD). That is, the difficulty of some items remained stable across different populations of students while some item difficulties changed. This suggests that we should expect difficulty drift over time and must constantly re-analyze and link back to a baseline set of data if we want to be able to use a stable scale (Rudner, 1983). This re-analysis is not discussed nor of concern to most users of indexed item banks. It is suspected that this phenomenon is not commonly understood nor is there a process which could be put in place to accommodate test developers which would be cost effective and simple to use.

In sum, item response theory proposes that item difficulty can be defined at a point in time, that difficulty is constant across populations and will remain constant over time. Adapting tests to an individual's ability is a matter of selecting items which match that individual's capabilities and given the tools available constructing tests is relatively painless.

Re-analysis of test data where the items are held constant but the examinee population is varied suggests that item difficulty is not as stable as purported for all items on a given instrument and in all contexts. This item drift has been noted both in a mathematical as well as an applied sense. In order to account for this item drift, a statistical link back to baseline data must be performed. Linking back to baseline with every test administration may represent a level of complexity which is well beyond the casual test developer's capabilities. Ironically, it was this group of individuals that these techniques were supposed to directly support. From an operational standpoint this suggests a commitment of staff, time and resources are necessary to maintain and operate a testing program based on IRT that may be beyond the grasp of most.

STANDARDIZED ACHIEVEMENT TESTS

Instruments which are prepared by publishing companies tend to follow statistical approaches which are more traditional. These instruments are often criticized for not meeting the exact curricular approaches used within a particular school district. Comments such as "...tests of this ilk measure broad areas but not our particular curriculum..." are commonplace. These statements indicate that more precise measurement is somehow needed and desirable.

It is certainly factual that publishing company tests must cover a very broad spectrum of curriculum. It is also the case that this spectrum is probably contained in one form or another in virtually every classroom. This suggests that there are also large segments of the curriculum which, are not tested with these instruments. If Bracey (1986) is correct in his analysis, and there is no better than a 50% match between standardized tests and curriculum, placing heavy reliance on point-in-time assessments will produce minimal usable information. Perhaps the debate regarding match and precision of measurement, IRT vs. standardized tests, is of far less concern than long-term need for additional categories of data.

In sum, publisher tests measure everyone's curriculum and at the same time no ones specific adaptations. They are broad in scope and most likely cover the curriculum elements most individuals define as \*basic skills. They produce a data set which will aggregate from an individual item response to a district wide summary of performance. Perhaps their most significant difference in contrast to IRT based tests is the availability of a contrast population as opposed to a difficulty index. The question for the user of these instruments must therefore be one of minor advantages. Selecting the CTBS U/V, the Portland Area Levels Test, CAT or SBS should probably be based more on the quality of reports and the degree of test/curriculum match than the item analytic and scale theory used to generate the test. Even more important is the relative contributions of these data within a larger information framework.

#### CRITERION REFERENCED - LOCALLY DEVELOPED

With the availability of item banks, many individuals have chosen the path of developing local tests. These individuals typically will organize a committee of teachers whose charge is first to define the curriculum which is in place. Using item banks, these individuals are asked to select items which match these objectives. Most frequently coverage of all goals is not available through these item banks. In this case these individuals are asked to develop specific items to cover the materials. Sometimes training on item writing is provided, some times not. Once a test has been constructed, typed, proofed and reproduced the committee disbands believing their job complete.

In some instances a pilot ' administration of the developed instrument is possible and practical. Most often the results of these pilot administrations reveal items which do not discriminate, based on IRT or more traditional item analysis techniques. This cycle in some test development efforts is repeated until such time that statistical analysis indicates that all included items do discriminate. Experience in these projects indicates that a

minimum of three to four months is needed to complete this cycle. This is a considerably different result than drawing items from a pool and having a test ready to be used.

An interesting question comes to mind in thinking through the potential effects of deleting items. If we repeatedly delete items from tests which are not meeting mathematical model expectations, over the course of time we may be producing a test which is free of curriculum effects. Perhaps following this approach in an attempt to produce a "tight" metric serves the purpose of retaining items which are in fact not impacted by those modifications to curricula which we continually strive to make. This may serve to insure that the resultant instrument is no more or less sensitive to curriculum than one published by a publishing house.

Individuals who participate in these experiences do appreciate not only the difficulty of test development but also that of defining a measurable unit of the curriculum. Tests are refined and sharpened to a level of specificity far removed from that which guided the initial project. Once these instruments are constructed they also tend to remain in place for extended periods of time. The idea of bringing a committee back together to redo a test to reflect a change to the adopted series is not frequently discussed. Using batteries of ill-maintained item banks and tests leaves one wondering about the magnitude of problems which are being created at the local district level.

In sum, producing CRT's from existing item banks is not a simple process. It does consume enormous energy and time and produces a result which over time may not be desirable. Integrating these instruments into a comprehensive testing program is often confusing in terms of defining a niche which seems most appropriate. Given the drift of item difficulty as well as the difficulty of maintaining tests, perhaps infrequent test developers are more appropriately advised to refrain from developing local instruments based on difficulty indexed items. At least item banking should not be embraced as a panacea for more precision or better curricular fit in educational measurement.

## INFORMATION SYSTEMS

What I have pointed out above is that item banks and associated metric approaches, in similar fashion to other metric approaches, have shortcomings. If used in a supervised fashion they are capable of producing usable data. Instruments produced by publishing companies also produce a level of usable data. The instruments developed at a local level also have a place in the data scheme of things. It is important not to become so tied to a particular methodology, however, that data options begin to be dictated by the method. Another way of saying this is the method becomes the objective.

From a practitioner's perspective the purported accuracy and precision of measurement gained from a particular approach loses itself in the realities of daily school operation. Data collected at points in time are valuable as describing that point in time. The particular measurement approach is not the key issue. The utility of any data point is from a broader perspective, key to the energy dedicated to the collection of those data.

While each field needs its champions, measurement being no exception, perhaps some tempering of technique with operational reality needs to occur.

The thrust on "school effectiveness" has begun to focus our attention on the notion of multiple data points collected over extended periods of time. The scores which we generate from a particular test instrument take a place juxtaposed with grade histories, attendance patterns, changes in family status and so on. This is beginning to suggest that as more data become available to the emphasis should begin to shift from a particular measurement technique to the development of schemes to examine data drawn from various sources reflecting a variety of perspectives.

Focusing energy on the organization and recombination of existing data appears to, at minimum, have promise for a large return in usable information. Facing the limitations of any testing technique or program squarely seems most appropriate. Recognizing limitations of data collection processes and moving forward with additional solutions to information problems promises many new frontiers for the measurement community.

#### SUMMARY

I have attempted to point out that no single measurement strategy will serve all purposes. The idea that we can examine an item at a single point in time with a sample of convenience and have that item remain constant in perpetuity seems to not be the case. With some individuals believing that accurate measurement of no more than 50% of the curriculum will occur, no matter what the instrument, then expending more and more energy on developing a "more accurate metrics seems to be time not well spent.

Investing in methods and techniques which allow a variety of data elements to be retrieved and juxtaposed may be a far better investment in the future. If we are able to recombine diverse data points to address a specific question regarding a trend, a student, a building and so on, we may be more closely responding to the goal of providing a better managed educational environment for the students whom we serve.

## REFERENCES

- Birnbaum, A., Some Latent Trait Models and Their Use in Inferring an Examinee's Abilities. In Lord, F.M. & Novick, M.R., Statistical Theory of Mental Test Scores. Addison-Wesley, 1968.
- Bracey, G.W., Mismatch of Testing and Instruction is Pervasive. Phi Delta Kappan, March, 1986.
- Lord, F.M., & Novick, M.R., Statistical Theory of Mental Test Scores Addison-Wesley, 1968.
- Occasional Papers, Multnomah Education Service District Measurement and Evaluation Department.
- Occasional Papers, Portland Public Schools, Departments of Research and Evaluation.
- Rudner, L.M., A Closer-Look At Latent Trait Parameter Invariance Educational and Psychological Measurement. 43. 1983.

## DISCUSSION

William E. Coffman University of Iowa

In reviewing the papers for this symposium, I was reminded of an experience I had some thirty years ago at a session on test theory at ETS with Dr. Fred Lord. At that time Dr. Lord was already deeply involved in the development of item response theory, but most of the test development at ETS was taking place in the context of classical test theory. I recall that I found myself deeply concerned by a consciousness of how inadequate classical test theory was. The theory assumed that any test consisted of a random sample of a population of test items and that the norms were based on a random sample of a population of individuals to be measured by the test, and it was quite clear, particularly in the case of test items, that we weren't dealing with random samples. I expressed my concern and Lord replied, "Yes, it's true that the model is an over-simplification of what we actually do, but it's really surprising how well the model usually works." As we have gained experience over the years, we have learned a great deal about the possibilities and limitations of the classical model, and competing theories have been developed that promise to overcome some of those limitations.

The symposium today is concerned with one of these competing test theories, one that is also based on an oversimplified model, and again we are told that it really works pretty well, even when the assumptions are not met very well. Our speakers tell us that the model overcomes some of the inadequacies of the classical model we have used for so long. The purpose of the symposium, according to Dean Forbes, was to introduce test directors and other persons involved in the operation of public school testing programs to the basic concepts involved in the so-called Rasch model. It was felt that many potential users are intimidated by Rasch and consider it too esoteric and abstract either for use or understanding. As a result, Forbes was given the charge to organize a symposium that would de-emphasize the technical complexities and let the audience see how I.R.T. can be used in practical educational measurement.

Let's try to summarize what the speakers have been telling us. Forbes has told us that the Rasch model assumes that, for a test that fits the model, a test score is the result of only two parameters, the difficulty of the items in the test and the ability of the test taker, and that both ability and difficulty are independent of the sample of individuals who may have taken the test and of the sample of items that are in a particular test. Also, that once the item difficulties have been determined, one can estimate the ability of any test taker very efficiently by simply administering items that are close in difficulty to the ability of the test taker. Furthermore, since the ability and

difficulty indices are sample free, once a bank of scaled items is created, the task of building a test is very simple: just select items of appropriate difficulty in relation to the ability of the examinees. Finally, since computer programs are available to do the job of providing the estimates, one doesn't need to understand the complexities of the theory in order to make use of the applications.

Hunt goes a step further. Given that the item bank fits the model, one is free to select from the bank only those items that are judged to be relevant to the local curriculum. Furthermore, "cross linking with national norms may be developed for a broader comparison." She notes the problem that current item banks deal only with basic skills areas and that there is need for item banks in the content fields such as science and social studies. Neither Forbes nor Hunt provide any data to support their arguments, but Forbes does cite references to the work of others.

Mershon reports on the experience of a school system in the development of criterion referenced tests and how the local item pool was expanded by relating it to more comprehensive item pools so that local items could use the scales that underlie the larger pools and so that items from the pools might be used for local tests. The paper introduces a note of caution: that for any particular goal, a test score should be based on a minimum of 20 items if one wishes to obtain acceptable reliability of the score. He does not, however, examine the question of how items that measure a variety of different goals can be scaled together in a common pool.

In the final paper, Thomas raises a number of questions about a tendency to accept claims for the model without question. He reports that item difficulty indices are not typically as stable as the Rasch theory would have us believe. Item difficulty indices do turn out to be different when they are calibrated on samples of different ability. Also, they differ from time to time, even for comparable samples of test takers. And since the item statistics for a pool can "drift", it's probably a good idea to have as an additional reference point data from a nationally normed test. Then too, since locally developed test items may vary in quality, it is desirable to be sure that they do fit the model well enough so that one can have some confidence in the scaling. Those items that do not fit the model need to be discarded. However, he continues, "if we repeatedly delete items from tests which are not meeting mathematical model expectations, over the course of time we may be producing a test which is free of curricular-effects . Per following this approach in an attempt to produce a 'tight' metric serves the purpose of retaining items which are in fact not impacted by those modifications to curricula which we continually strive to make."

One sentence in Thomas' paper warrants special attention. He writes, "the difficulty of some items remained stable across different populations (*italics mine*) of students while some item difficulties changed." Here Thomas has put his finger on a critical issue in evaluating the appropriateness of basing testing practices on the Rasch model. The theory is that when the data fit the model, item difficulty indices and person ability estimates are sample free, not population free. This means that one who proposes to use the Rasch model needs to think very clearly about whether or not the individual for whom the test is intended belong to the population on which the items were scaled -- or whether the particular set of items in the test are still measuring the same ability for these individuals that they were measuring for the individual. in the scaling group.

Those who are most enthusiastic about the use of the Rasch model assure us that the model is robust and that one doesn't need to be concerned about the fact that the model is an oversimplification of the situation in the real world of educational tests. The computer can take care of the mathematical details. All you have to do is trust the programs to produce numbers that tell you what you want to know. But are there dangers one needs to be aware of?

Many years before the Rasch model was proposed, Ledyard Tucker was already studying the appropriateness of item characteristics curves as describers of test questions and was conducting research on the scalability of test items. He reasoned that English vocabulary items were likely to be scalable if any were, but he was also aware that he ought to have a clear idea of whether or not they would behave in the same way for different subsets of the student population. He therefore carried out the scaling on four different groups: (1) high socioeconomic northeasters, (2) high socioeconomic southeasters, (3) low socioeconomic northeasters, and (4) low socioeconomic southeasters. He concluded that most vocabulary items would scale, that is, that they had the same relative difficulty from group to group. For most of the items, the relative difficulty was 1, 2, 3, 4 as had been predicted. But there were exceptions. For example, the relative difficulty of the word "scorpion" was 2,1,3,4. Scorpions are more familiar to southeasters, but the low socioeconomic southeasters call them "sting lizards." As any student of the Rasch model knows, items that don't fit the model are to be kept out of the pool. It is questionable, however, whether the experimental procedures used to determine fit are as appropriate as those used by Tucker many years ago.

One might raise the question, for example, of whether or not the same scale values would be obtained if one were to take 10% of the vocabulary items from a pool, drill the students in a

particular school on these vocabulary words only for three weeks, and then scale the items again using only those students. There is also the question of the effects of the passage of time on the difficulty of certain test items. I recall that we found for items in the College Board Social Studies pool dealing with presidential elections the p-values varied in four-year cycles corresponding to the years of presidential elections. I believe a similar pattern would be found with Rasch difficulty indices.

These problems are well understood by those who deal with test theory. For example, I found this statement on page 66 of *Best Test Design* by Wright and Stone: Even if the measurement model tends to fit a particular application, we cannot predict in advance how well new items (or even old ones) will continue to work in every situation in which they might be applied, nor can we know in advance how all persons will always respond. Therefore, if we are serious in our attempt to measure, we must examine every application to see how well each set of responses corresponds to our model expectations. We must evaluate not only the plausibility of the sample personal responses, but also the plausibility of each person's responses to the set of items in his test. To do this we must examine the response of each person to each item to determine whether it is consistent with the general pattern of responses observed. I see little evidence that those who are enthusiastic about applications of the Rasch model are following this recommendation.

To those of you who are considering the use of items from item pools that are scaled using the Rasch model, let me urge these cautions:

- 1) When the individuals being measured are a part of the population that manifests the ability defined by the pool of items that defines the ability, both ability and difficulty are sample free; but the question of fit to the model is an empirical question to be established and not a theoretical answer to be assumed.

- 2) Often, in applying the Rasch Model to develop a comprehensive scale for an item pool, the assumption is made that there is an approximately normal distribution of ability in the population and an approximately normal distribution of difficulty for the item pool. Normal distributions are most likely to occur as a result of the operation of a large variety of elements over a long period of time--as in the development of general language ability. Any limited number of systematic elements, as in the application of a carefully organized teaching effort, is likely to produce departures from the model. In fact, one might argue that the purpose of a school curriculum is to produce learning that is non-random and therefore learning that is distributed in other than a

normal pattern. -I suspect that the more we learn about how to teach, the less likely it is that achievement test items will fit the Rasch model, that is, unless we divide the item pool into a large number of independent sets, one for each specific objective.

3) Recall that the Rasch model is based on the same basic information as the classical model, that is, the responses of individuals to test items; and there is nothing magical about the numbers that are produced by application of the model. You still have to think about what interpretations are legitimate for the numbers generated by your application of the model. Major advantages of the model are that it permits one to focus measurement at particular regions of the ability scale, that it is possible to assess the efficiency of the measurement at different points on the scale, and that an analysis of responses can tell you when things are going wrong. Unfortunately, in order to know what to pay attention to in the analysis, you have to understand the model, not simply to let the computer produce numbers to be interpreted automatically.

In this context, it is interesting to note that a Rasch analysis has been used to identify items that are biased against certain minority groups. Notice, however, that when such items are identified, it is not always appropriate simply to drop the discrepant items from the test, assuming that the scaling of the remaining items is correct. It depends on how much the items that are dropped contributed to the original scaling. Of course, there is always the possibility that if you rescale the remaining items, the scale value won't change very-much. Certainly this is likely to be the case if there are only a few items identified as biased in a large pool of items. If however, one is choosing only a limited number of items from a pool that are identified as appropriate because they are judged to measure those objectives taught in a particular curriculum, then those items may actually be defining a different ability from that represented by the original pool of items.

4) Finally, if you propose to make use of the Rasch model in generating your test scores, recognize that the better you understand the model and its implications, the more likely you are to make sensible interpretations of the test scores you generate. Don't be mesmerized by the sales pitches of those who are marketing item pools or computer programs. Insist on understanding what it is you are buying before you close the deal.