Measure Accuracy:
Functioning-Level vs. Grade-Level Testing

George S. Ingebo, Ph.D.

In grade-level testing, all grade-three students take the same grade-three test; grade-four students the grade-four test.

In functioning-level testing, students take tests designed for their attainment level, whether they are low-middle or high-achieving students. There may be five to six achievement levels at a grade-level level.

Functioning-level testing is not equivalent to out-of-level testing, in which low- or high-achieving students are tested at a lower or higher grade-level.

Data from test publishers' grade-level tests indicate that few grade-level test items accurately measure low- or high-achieving students' ability. This study shows that, when compared to testing students at grade-level, testing students at their functioning-level substantially reduces measurement error.

Since 1979 to the present, the Portland Public Schools of Portland, Oregon administers a basic skills testing system using functioning-level tests. Portland calibrates this testing system with a Rasch measurement model and maintains records on the performance of the students taking these tests at every grade during that entire time. Students take one achievement test out of a series of tests in the fall, and another in the spring.

Portland Public Schools selects a test level for a student by finding each student's score on their last district test. Portland places a student at an ability level based on their expected growth. By fitting a test to a student's established ability, whether it is a high- or low-achieving student, sufficient items in each functioning level test measure the performance of that student.

The State of Oregon State Assessment Program tests at grade-level rather than functioning-level. This study compares the Portland Public Schools functional-level testing with the State of Oregon grade-level testing results.

The State of Oregon employs the same testing procedures and the same Rasch scale used in the Portland Public Schools districts. The Oregon State and the Portland Public Schools have the same curricular goals in reading and mathematics. Both Portland Public Schools and Oregon State testing systems calibrate their tests for content difficulty with the same Rasch scaling model. These factors facilitate direct comparisons.

The State of Oregon administers grade-level tests to students once a year, in the spring. There are two state tests for mathematics and two for reading. Depending on where they live, students take one each of these state tests. The Portland Public Schools administers level-tests in fall and spring.

Procedure

We generate a quantitative probability that, given a test scaled in calibrated measures, each item in that test has a level of difficulty based on this scale. We predict the percentage of students who we expect to answer each item correctly from the calibrated item measures.

> *We expect that 50 percent of students who have attained an ability measure of 200 on previous tests to correctly answer an item that has a difficulty level calibrated at 200. We expect 25 percent of students with an ability level of 190 to correctly answer an item with a difficulty level calibrated at 200. We expect 75 percent of students with an ability level of 210 to correctly answer an item with a difficulty level calibrated at 200.*

Then, we estimate differences between expected and actual performance of every item of each student group achieving the same Rasch scale total score. We group students from low-, middle-, or high-achieving, based on their past ability measures.

We define test accuracy based on the amount, not the number, of deviations from the expected score. Tests with greater deviation amounts are less accurate. Tests with the deviation closer to expectation are more accurate.

Grade-level tests data were from 1993-94 state grade-five mathematics tests administered to two groups of students and another two groups of students taking grade-five reading tests.

Functioning-level tests data were from only one of five 1993 Portland Schools levels-tests administered to grade-five students.

We compute Rasch scale measures for all Oregon State

TEACHING

Assessment grade-five 1993-94 tests and the Portland spring 1994 functioning-level test scores. We exclude records of students getting less than 30 percent items correct from the analysis.

For each student scale score, we recover the probability of success for that student on each item in the test taken.

For all students getting the same Rasch scale measure in each compared group, we compute the differences between expected and actual performance.

We examine the differences between mean expected scores and the mean actual scores on all test items completed by each total-score group.

We examine the differences between the expected standard deviation and the actual standard deviation on all items attempted by students at each raw score level.

We aggregate the differences between expected and actual performance all items for all tests (reading and math) per increasing student measures.

Findings

1. For Oregon State grade-level tests and the Portland functioning-level tests, students had the same rate of number of differences. In both student groups, a similar ratio did better or worse than expected, Table 1.

2. The amount of difference between expected and actual performance is twice to three times as great for low-achieving reading students taking State of Oregon grade-level tests, Figures 1b and c, as for those taking the Portland schools functioning-level tests, Figure 1a.

3. The difference between expected and actual performance for low-achieving mathematics students is twice to three times more on the State of Oregon grade-level tests, Figures 1e and f, than the Portland functioning-level tests, Figure 1d.

4. Reading and math standard deviations aggregates show functioning-level tests are two times more accurate than grade-level tests, Figure 2.

5. Basic skills measures for all students are best for students with mid-range scores for grade-level and functioning-level tests. Differences in measurement accuracy between the grade-level and the functioning-level groups are less pronounced at the upper score levels than at the low score levels.

a) Oregon State Grade-level Tests

Grade-level Tests		Number of Students	Number of Differences	Items
Reading	351	7,545	1,064	38
	352	7,643	1,044	36
Math	451	7,443	1,380	46
	452	7,347	1,334	46
total		29,978	4,822	166

vs. b) Portland Area Functioning-Level Tests

Functioning-level Tests		Number of Students	Number of Differences	Items
Reading	504	7,512	1,000	40
Math	516	13,272	2,220	60
total		20,585	3,220	100

Table 1

Conclusions

Students grade-level tests have unacceptable measurement error, especially with low-achieving students. The functioning-level test measures are two to three times more accurate than the grade-level test scores for predicting low-achieving students' achievement. This raises concern over the continued use of grade-level tests for student placement and school program evaluation.

When used with the same students, functioning-level tests like those used in the Portland Schools give more accurate assessments than the grade-level tests.

Functioning-level tests using item banks in which all items are calibrated to a single scale of difficulty accurately test students from the lowest grade-three level to the highest grade-eight level, Figure 3.

Figures 1, 2, and 3 are on next page (68).

George S. Ingebo, Ph.D.

Dr. Ingebo brings a wealth of life experience to student achievement testing. He grew up in Winnett, Montana, a small town in Petroleum County, Montana, where he learned to box at Shorty's Gym. During WWII, he flew combat missions with a B-24 bomber crew in India and China. He earned a Ph.D. from the University of Washington-Seattle. He taught high school science and mathematics, coached football and track.

A pioneer in constructing standardized machine scored tests, Dr. Ingebo established a college entrance testing program and developed predictors for college success. He directed a child clinical testing service at the University of Pacific. In 1969 Dr. Ingebo switched directions in testing. After hearing a lecture on Rasch Model testing by Ben Wright, Dr. Ingebo introduced this model into the Portland Schools. He established a new school testing program in Portland Elementary Schools. He provided technical planning in the Portland Metropolitan Area School Districts' High school testing cooperative. He helped found the Metropolitan Districts' Northwest Evaluation Association. He developed a variety of techniques for program evaluations based on Rasch equal interval measures from levels tests. He conducted research on the use of the Rasch Model over a 16-year period. Following on the Portland success, the Rasch model is increasingly used for measuring student achievement in metropolitan school districts.

Standard Deviation of Differences Between Expected and Actual Student Performance

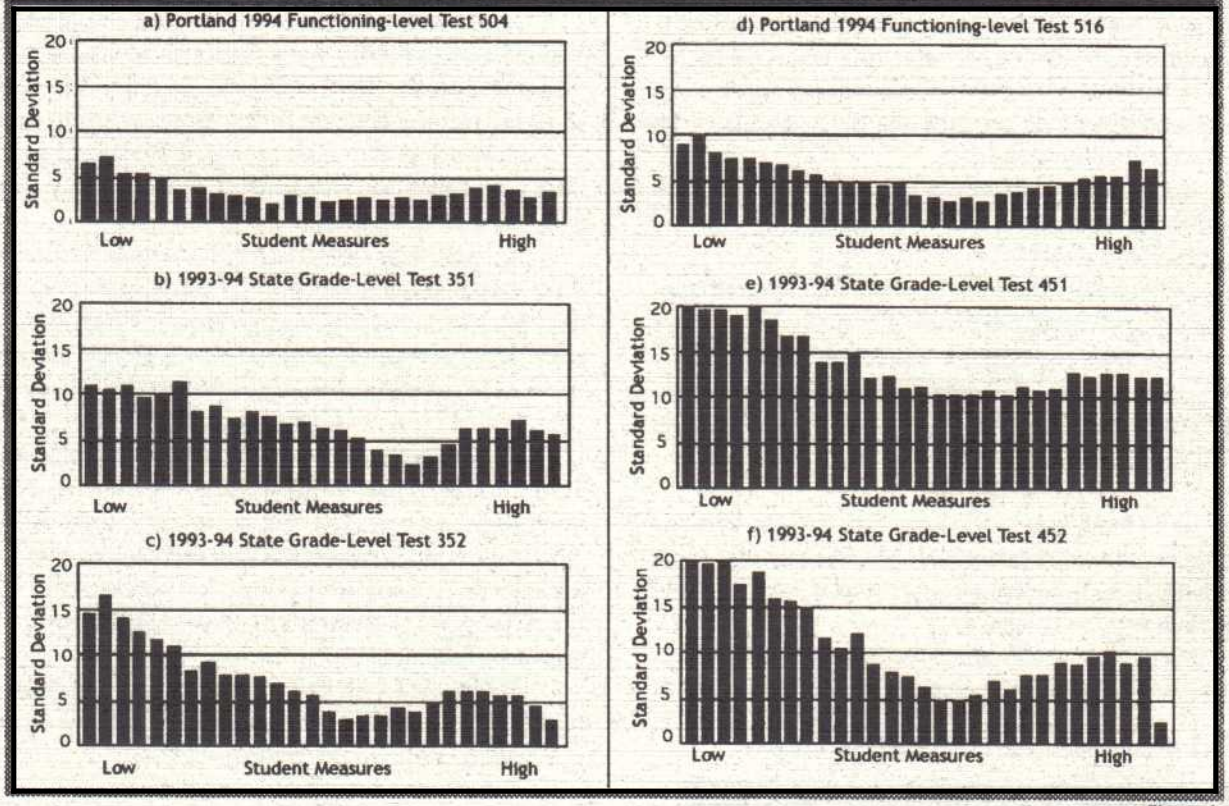
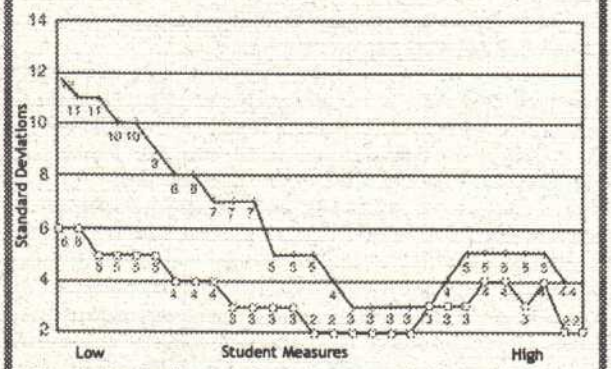


Figure 1

Plots of Aggregated Standard Deviation Differences of Expected and Actual Performance for All Students Taking 1993-94 Grade-Level and 1994 Functioning-Level Tests

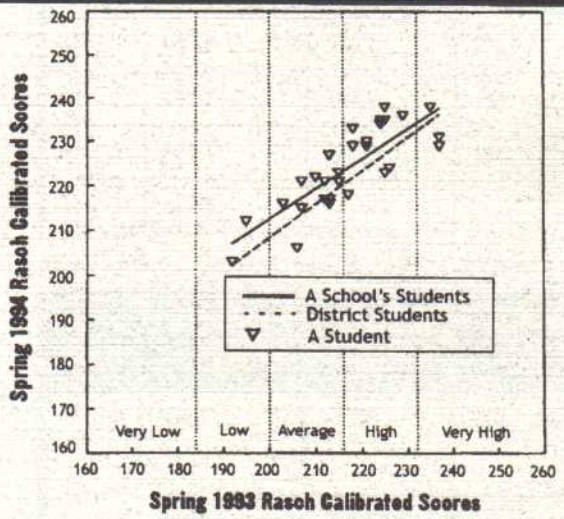
Reading and Mathematics Groups Combined



Functioning-level test accuracy improves with the middle and upper level students.

Figure 2

1994 Portland Public Schools Student Gains Compared to 1993 Test Scores



Comparing student triangles to the reference lines shows whether their gain is greater (above the line) or less (below the line). Comparing the two reference lines shows whether, on the average, students in a school gain more or less than similar level students in the rest of the District

Figure 3