UNIVERSITY OF DELAWARE

EDUCATION RESEARCH & DEVELOPMENT CENTER

Education Policy Brief

VOLUME 16, MAY 2004

Testing: Not an Exact Science

"The public place great faith in the infallibility of test results."

National Board on Educational Testing and Public Policy

In Delaware, as most states, student tests are the foundation for accountability systems. As federal legislation, such as *No Child Left Behind*, has sharpened the focus on student achievement, state testing practices have become increasingly consequential for students, teachers, and school systems. Despite the appearance of mathematical exactness in a numerical score, standardized achievement tests do not yield exact measurements of what individuals know and can do. Even tests that are well designed and properly administered are inevitably subject to both statistical and human error.[i] The intent of this Education Policy Brief is to acquaint Delaware policymakers with some fundamental concepts[*] about testing and the interpretation of test results. A full appreciation of these concepts is critical to sound policymaking as the state proceeds with its educational accountability agenda.

For more information or questions regarding this Education Policy Brief, contact:

Audrey J. Noble, Ph.D., Director
Delaware Education Research & Development Center
Phone: 302-831-4433
E-mail: ajnoble@udel.edu

With contributions from Dr. Robert Mislevy and Karen Douglas, University of Maryland and
from Dr. Ratna Nandkumar and Heidi Sweetman, University of Delaware

[*] A key source of information that informs this policy brief comes from a report developed by the National Research Council entitled: *High Stakes: Testing for Tracking, Promotion, and Graduation.* This book, written in 1999, is the result of the work of the nation's most prominent measurement experts and their advice to the Clinton administration about its plan to create a voluntary national test. Other sources are cited throughout; full references can be found in the extended length document at www.rdc.udel.edu

TESTING INCLUDES ERROR.

Educational tests are subject to far more limitations in their accuracy than many people realize. These limitations are inherent, in varying degrees, to all educational assessments. Educational testing is more a process of careful estimation than one of precise measurement. Consequently, an individual student's test score should be seen as a rough approximation [ii]of one's performance in a particular setting, not an exact assessment of one's ability.

A measurement concept that addresses the inherent variability of test results is *reliability.* It refers to the stability or reproducibility of a test's results.[iii] One can think of reliability in terms of the likelihood that a student's score would change if he took an equivalent test the next day. Experts say that if test scores are used to make high-stakes decisions about individual students, such as promotion or graduation, it is imperative that the scores be stable indicators of performance.[iv] Since tests are not perfect, a student's score can be expected to vary across the different versions of a test – within a margin of error determined by the reliability of the test. Reliability is an important concept when comparing different tests.

To compare scores from the *same* test, a more useful index than reliability is the *standard error of measurement* (SEM), which reflects the *unreliability* of a test.[v] The SEM defines a range of likely variation or uncertainty around a test score, similar to the margin of error of +/- points used in reporting polling results. Defining the margin of error around a specific score "reminds us that scores earned by students on commercial (or classroom) tests are not exact."[vi]

Some factors that influence the reliability and SEM of assessments have nothing to do with instruction. They include the length of the test, how it is scored, and the clarity of its questions. The Standards for Education Accountability Systems issued by the National Center for Research on Evaluation, Standards, and Student Testing indicate that "if test data are used as a basis of rewards and sanctions, evidence of technical quality of the measures and error rates associated with misclassification of individuals or institutions should be published."[vii] In other words, the impact of the reliability and the SEM of a test needs to be fully disclosed when test results are used to make high-stakes decisions about students or schools.

STANDARD SETTING PROCEDURES ARE SUBJECTIVE.

Cut scores are the points on a scale of test scores that designate levels of performance from excellent to acceptable to unacceptable. Even when done with care and in a principled manner, designating cut scores is another potential source of human error in educational testing. The methods that are used to set cut scores all rely on some sort of potentially fallible human judgment.[viii] According to the National Board on Educational Testing and Public Policy, "performance levels are based on cut scores. Cut scores, in turn, are based on judgment. The problem is, as long as there is judgment involved in the cut-score setting procedure, we can never be completely sure performance levels accurately reflect student achievement."[ix]

IMPORTANT DECISIONS SHOULD NOT BE BASED ON A SINGLE TEST SCORE.

Tests and standard-setting procedures involve both human and statistical error. For these reasons, numerous professional and educational research organizations[x] warn that high-stakes decisions should not be based on a single test score. They state that "no single test score can be considered a definitive measure of a student's knowledge."[xi] "Scores from large-scale assessments should never be the only source of information to make promotion or retention decisions." "Decisions that affect individual students' life chances or educational opportunities should not be made on the basis of test scores alone."[xii] The nature of educational measurement makes any test vulnerable to error and important decisions should never be made based on a single assessment.

THE DELAWARE SITUATION

Reliability of the Delaware Student Testing Program[xiii]

"A satisfactory level of reliability depends on how a measure is being used… A great deal hinges on exact test scores when decisions are made about individuals…If important decisions are made with respect to specific test scores, a reliability of .90 is the bare minimum, and a reliability of .95 should be considered the desirable standard."[xiv] The table below illustrates the reliability and standard errors of measurement of the DSTP reading and math assessments from 1998 to 2002.

Table 1. 1998-2002 DSTP Reading Reliability Coefficients and Standard Errors of Measurement

	1998		1999		2000		2001		2002	
	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM
Grade 3	.93	11.6	.92	11.7	.90	12.2	.91	11.4	.91	11.3
Grade 5	.94	11.7	.93	11.5	.91	12.3	.92	11.9	.90	11.7
Grade 8	.92	11.9	.92	11.7	.91	11.7	.91	11.2	.90	11.2
Grade 10	.92	12.2	.92	12.2	.91	12.5	.92	11.4	.91	11.4

Table 2. 1998-2002 DSTP Math Reliability Coefficients and Standard Errors of Measurement

	1998		1999		2000		2001		2002	
	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM	Rel.	SEM
Grade 3	.91	12.2	.91	12.6	.90	12.6	.91	12.7	.92	12.3
Grade 5	.92	11.8	.92	12.1	.91	11.7	.92	11.3	.92	11.0
Grade 8	.93	11.9	.91	11.6	.92	11.5	.92	11.0	.92	11.0
Grade 10	.92	11.2	.91	11.3	.92	11.0	.92	10.6	.92	11.0

Between 1998 and 2002 the reliability of the Delaware Student Testing Program never reached the "desirable standard" of .95 and at times has only reached the "bare minimum" of .90. Nonetheless, during this time period important decisions about individual students have been made based on this assessment including retention in grade, attendance in summer school, placement in remedial programs, and qualification for scholarships.

Probability of DSTP Performance Level Misclassification

As reviewed above, the reliability of a test provides an important, but partial, picture of the quality of a test and the trustworthiness of its results. It is also important to look at the SEM especially when considering the accuracy of assigning students to various performance levels based on their test scores. "Measurement error that is associated with any test score results in classification errors…Valid inferences about student proficiency are undermined by measurement errors that result in misclassification of students. Hence, it is critical that the probability of misclassification be evaluated."[xv]

To do this, Dr. Robert Mislevy, a psychometric expert from the University of Maryland, conducted a simulation study based on the 2001 DSTP third grade reading scores.[xvi] His study examined the reliability and SEM of DSTP scaled scores to determine how often students were misclassified on the state's performance level scale. Mislevy found that in 2001 77% of third grade students were accurately classified in reading. Consequently, in 2001 some 23% of Delaware's 3rd grade students were misclassified.

Mislevy states that decision errors are the “inevitable consequence” of imperfect measurement. These errors could have resulted in a child’s being inappropriately retained in grade, or unnecessarily required to attend summer school, or incorrectly placed in a special reading program and subsequently denied other learning opportunities. In addition to educational impact, there are other potential effects of misclassification that results in retention. There is research that indicates that many children feel stigmatized by grade retention. One study showed that students rank grade retention as the third most feared life experience behind blindness and the death of a parent. [xvii]

Researchers at the University of Delaware replicated Mislevy’s analysis using the 2003 DSTP eighth grade mathematics scores to examine potential misclassification at that level. [xviii] In this case, it was found that 75% of 8th grade students who took the math DSTP in 2003 were accurately classified, leaving 25% suffering from the “inevitable consequence” of imperfect measurement. This misclassification could have resulted in the same consequences that may have resulted for third grade students discussed earlier. It is also important to recognize that students’ 8th grade mathematics performance has particular significance in regards to their ability to enroll in higher levels of mathematics in high school and subsequent access to higher education.

Confidence in Cut Score Setting Process

The American Education Research Association states in its Standards[xix] that “whenever cut scores are used, the quality of the standard-setting process should be documented and evaluated- including the qualification of the judges, the method or methods employed, and the degree of consensus reached.” In 1999, Delaware DOE released a report[xx] which addressed this issue. In regards to the degree of consensus reached, feedback[xxi] from judges involved in setting Delaware’s cut scores indicated that 27% of the judges were “highly confident” that the description of the PL4 (exceeds the standard) cut point was reasonable; 20% were “highly confident” that the description of the PL3 (meets the standard) cut point was reasonable.

WHAT DO THE EXPERTS RECOMMEND?

“Blanket criticisms of testing and assessment are not justified. When tests are used in ways that meet relevant psychometric, legal, and educational standards, students’ scores provide important information that, combined with information from other sources, can lead to decisions that promote student learning….It is also a mistake to accept observed test scores as either infallible or immutable.”[xxii]

National experts[xxiii] maintain that “performance-based accountability systems are, to say the least, works in progress… Tests on which stakes are based are fallible and limited measures; the statements they make about student and school performance carry margins of error for both students and schools, making clear judgments about performance difficult. These limits of tests are overlooked routinely in current accountability policies.”[xxiv]

Back in 1997, Delaware’s Business/Public Education Council released its well-known report, *The Missing Link.* Among its recommendations about accountability, the Council urged employing “multiple ways to assess performance.”[xxv] Their recommendations parallel current experts’ [xxvi] antidotes for test misuse: strong curriculum-embedded assessments, knowledgeable use of alternative assessments, multiple measures of instructional quality and student performance with no high stakes decisions based on any single measure.

ENDNOTES

[i] *High Stakes: Testing for Tracking, Promotion, and Graduation, p. 14.*

[ii] W. James Popham, *America's Failing Schools, p55.*

[iii] Ibid, pp. 71-72.

[iv] Hamilton, L.S., Stecher, B.M, and Klein, S.P. (2002) *Making Sense of Test-Based Accountability in Education,* RAND, p. 52.

[v] *High Stakes: Testing for Tracking, Promotion, and Graduation, p. 72.*

[vi] W. James Popham, *Classroom Assessment: What Teachers Need to Know, p. 33*

[vii] CRESST, Standards for Education Accountability Systems, winter 2002, p.3.

[viii] *High Stakes: Testing for Tracking, Promotion, and Graduation, p. 99*

[ix] National Board on Educational Testing and Public Policy: *Cut Scores: Results May Vary,* Horn, c., Ramos,,M., Blumer, I., and Madaus, G., pp. 29-30.

[x] These professional groups include but are not limited to the National Research Council, the American Educational Research Association, the American Psychological Association, the National Parent Teachers Association, the National Education Association , the National Board on Educational Testing and Public Policy, RAND, the National Center for Research on Evaluation, Standards, and Student Testing, the National Council on Measurement in Education, the American Counseling Association, the American Speech-Language-Hearing Association, the National Association of School Psychologists, and the National Association of Test Directors.

[xi] *High Stakes: Testing for Tracking, Promotion, and Graduation, p. 3.*

[xii] AERA Position Statement Concerning High-Stakes Testing in PreK-12 Education, July 2000

[xiii] Data derived from 1998-2002 DSTP Technical reports, source http://www.doe.state.de.us/

[xiv] *Psychometric Theory,* Nunnally, J.C., & Bernstein, I.H. p. 264-265.

[xv] Baker & Linn in Fuhrman & Elmore, p. 55

[xvi] Mislevy & Douglas, "Given fallible measures, how often will we make the right (or wrong) classification based on the observed test score?" We conducted a simple simulation study to show another way of answering this question. In real data we never can know with certainty what the true scores are for students. In simulated studies, however, we stipulate "true" scores, and then study the distribution of the "observed scores" likely to result from these "true" scores given our model assumptions and approximations from published data. By extension, the relationships exhibited in this simulated approach should mirror those in real data.

The results are summarized below. The "bottom line" is that an estimated 77% of students are accurately classified on third grade reading scores from the DSTP. If a pass/fail decision were to be made in which the criterion was "meets standards," then approximately 93% would be accurately classified. Whether this is sufficient accuracy for a particular decision is a matter for thoughtful consideration, which is certainly facilitated by estimating the accuracy of the decision. In addition, the estimation of expected false positives and false negatives may also be important information for decision makers as each type of error has different implications for students and for schools. A false positive decision promotes a student who is did not really meet the criterion, whereas a false negative decision holds a student back who deserves to be promoted. The results below are based on your analysis, and are one way that the implications might be communicated to your readers, perhaps with a better feel for the consequences of the measurement errors.

Simulation Study: Potential for Misclassification on DSTP

This project is a simulation study to investigate the potential misclassification on the basis of third grade reading scores from Delaware State Testing Program (DSTP). The simulation is based on statewide results in 2001.

N = 8,394
Mean = 435.17
Standard Deviation = 38.61
Reliability = .91
Standard Error of Measurement = 11.4

DSTP classifies the performance of students on the third-grade reading test based on the following cut-scores:

Level	Category	Criterion
Well below standard	1	Below 387
Below standard	2	387 - 410
Meets standard	3	411 - 464
Exceeds standard	4	465 - 481
Distinguished	5	482 or higher

Data were simulated using WinBUGS, Version 1.4. WinBUGS uses Markov Chain Monte Carlo methods (MCMC) and Bayesian inference to simulate and solve complex estimation problems. (See http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml for more information on WinBUGS).

A sample of true scores was generated for 10,000 hypothetical students with a distribution mirroring that for third grade reading (as described above). An observed score was also generated for each true score by forming a conditional distribution with a mean equal to the true score and a standard deviation equal to the SEm. For each hypothetical student, one score was randomly selected from this conditional distribution. A two-way contingency table was formulated that indexed the classification of the true score by the classification of the accompanying observed score (see Table 1 below) for each hypothetical student. The simulation accurately captured the distribution of true scores, with a mean of 435.1 and standard deviation of 38.57. Observed scores also had a mean of 435.1, with a slightly higher standard deviation (40.23). This is to be expected – with less than perfect reliability, regression to the mean is expected when estimating true scores from observed scores. Therefore, there is more variability in the observed scores than the true scores.

Table 1 shows the accuracy of classification in the simulation study. Cells in bold (all cells on the diagonal) represent a consistent decision between the true score and observed score for each hypothetical student. All other cells represent inaccurate decisions.

Table 1: Classification of hypothetical students based on true score and observed score.

	OBSERVED					
TRUE	1	2	3	4	5	Total
1	**847**	153	3	0	0	1003
2	256	**994**	325	0	0	1575
3	8	424	**4328**	367	31	5158
4	0	0	292	**552**	254	1098
5	0	0	14	164	**988**	1166
Total	1111	1571	4962	1083	1273	10000

Overall, 77% of students were classified accurately (i.e., the same classification was made for true and observed score). Twelve percent received a lower observed classification than true classification (i.e., false negatives). Eleven percent received a higher observed classification than true classification (i.e., false positives).

Table 2 presents the same information in a different format. Each proportion represents the conditional probability of obtaining an accurate classification for each true score.
For example, for students whose true classification was a “1”, 84 out of 100 also had an observed score classification of “1.”

Table 2: Proportion of hypothetical students in each observed category based on true category.

	OBSERVED					
TRUE	1	2	3	4	5	Total
1	**0.84**	0.15	0.00	0.00	0.00	1.00
2	0.16	**0.63**	0.21	0.00	0.00	1.00
3	0.00	0.08	**0.84**	0.07	0.01	1.00
4	0.00	0.00	0.27	**0.50**	0.23	1.00
5	0.00	0.00	0.01	0.14	**0.85**	1.00
Total	1.01	0.87	1.33	0.71	1.08	5.00

Note: Proportions may not add to 1 due to rounding.

This table illustrates that categories 1, 3, and 5 had a higher proportion of accurate decisions, whereas categories 2 and 4 had much lower proportions. As you point out on pg. 8 of your report, this is a reflection of the variable widths of the score ranges for categories. Category 2 includes 23 possible scores (387-410), and Category 4 includes 16 (465-481), whereas Category 3 includes 53 possible scores (411-464). The more narrow the score range in a category, the greater the likelihood that a student’s true score distribution will overlap with a contiguous score category. The width of categories 1 and 5 cannot be determined based on available data, but they are also subject to floor and ceiling effects and are only affected by misclassification in one direction.

Tables 1 and 2 show the likelihood of misclassification based on all five categories. Many decisions about students apply a single cut-score. For example, a student may not be promoted to the next grade unless he or she meets an acceptable level of performance. Therefore, a single cut-score decision was applied to the simulated data to illustrate the likelihood of making an inaccurate decision in such a situation. The decision applied was whether the student reached category 3, which is described in the DSTP report as "meets standards."

Table 3 shows the number of students who received an accurate decision, and the same information is presented as proportions in Table 4. An accurate decision was made for approximately 93% of students. For those who received inaccurate decisions, 3% were false positives and 4% were false negatives.

Table 3: Number of hypothetical students passing "meets standards" criterion

	OBSERVED		
TRUE	fail	pass	Total
fail	**2250**	328	2578
pass	432	**6990**	7422
Total	2682	7318	10000

Table 4: Proportion of hypothetical students passing "meets standards" criterion

	OBSERVED		
TRUE	fail	pass	Total
fail	**0.23**	0.03	0.26
pass	0.04	**0.70**	0.74
Total	0.27	0.73	1.00

[xvii] Dill, S.V. (1993). Closing the Gap: Acceleration vs. Remediation and the Impact of Retention in Grade on Student Achievement. (ERIC Document Reproduction Services no. 354 938).
[xviii] Nandakumar, R. & Sweetman, H.

Using the methodology employed by Robert Mislevy and Karen Douglas, below is a simulation study to investigate the potential misclassification on the basis of eighth grade mathematics scores from Delaware State Testing Program (DSTP). The simulation is based on the statewide results in spring 2003.

2003 Mathematics DSTP Grade 8
N=9,468
Mean=493.98
Standard Deviation=38.97
Reliability=0.92
Standard Error of Measurement=11.0

DSTP classifies the performance of students in the eighth grade mathematics test based on the following cut-scores:

Level	Category	Criterion
Well below standard	1	Below 468
Below standard	2	469 to 492
Meets standard	3	493-530
Exceeds standard	4	531-548
Distinguished	5	549 or more

Again, the same methodology employed by Mislevy and Douglas was utilized to construct the contingency table below. Once again, the simulation accurately captured the distribution of the true score, with a mean of 494.1 and standard deviation of 38.88. The observed scores had a mean of 494.2, and once again, a slightly higher standard deviation of 40.22. As previously explained, this is to be expected with less than perfect reliability, regression to the mean is expected when estimating true scores from observed scores. The slightly larger standard deviation indicates that there is more variability in the observed scores than the true scores.

Table 1 shows the accuracy of classification in the simulation study. The cells in bold (all cells on the diagonal) represent a compatible decision between the true score and observed score for each hypothetical student. All other cells represent inaccurate decisions.

Table 1: Classification of hypothetical students based on true score and observed score.

	Observed					
True	**1**	**2**	**3**	**4**	**5**	**Total**
1	**2170**	313	8	0	0	2491
2	372	**1464**	441	0	0	2277
3	7	445	**2664**	328	16	3460
4	0	0	228	**518**	188	934
5	0	0	8	129	**701**	838
Total	2549	2222	3349	975	905	10000

Table 2 displays the same information as above, but in a slightly different format. As in the other study, each proportion represents the conditional probability of obtaining an accurate classification for each true score.

Table 2: Proportion of hypothetical students in each observed category based on true category.

	Observed					
True	**1**	**2**	**3**	**4**	**5**	**Total**
1	**0.871136**	0.125652	0.003212	0	0	1
2	0.163373	**0.642951**	0.193676	0	0	1
3	0.002023	0.128613	**0.769942**	0.094798	0.004624	1
4	0	0	0.244111	**0.554604**	0.201285	1
5	0	0	0.009547	0.153938	**0.836516**	1
Total	1.036532	0.897216	1.220488	0.803339	1.042425	5

The proportions indicate that categories 1, 3, and 5 had a higher proportion of accurate decisions, while categories 2 and 4 had lower proportions of accurate decisions. Once again, this is in part a function of the variable widths of the score ranges for each category. Category 2 includes 23 possible scores (469-492) and Category 4 includes 17 possible scores. The smaller score ranges of Categories 2 and 4 can be compared to the considerably larger score range of Category 3 (493-530) which includes 37 possible scores.

Tables 1 and 2 show the likelihood of misclassification based on all five categories. However, as noted by Mislevy and Douglas, many decision about students are made using a single cut-score.

Table 3 shows the number of students who received an accurate decision based on a single cut score. The same information is presented as proportions in Table 4. The data indicate that accurate decision was made for approximately 92% of students. For those who received inaccurate decisions, 4% were false positives and 4% were false negatives.

Table 3: Number of hypothetical students passing "meets standards" criterion.

Observed			
True	**Fail**	**Pass**	**Total**
Fail	**0.4382**	0.0432	0.4814
Pass	0.0426	**0.476**	0.5186
Total	0.4808	0.5192	1

Table 4: Proportion of hypothetical students passing "meets standards" criterion.

Observed			
True	**Fail**	**Pass**	**Total**
Fail	**4382**	432	4814
Pass	426	**4760**	5186
Total	4808	5192	10000

[xix] 1985, AERA Standard 6.9 1998

[xx] Delaware Department of Education (August, 1999) entitled *Establishing Proficiency Levels for the Delaware Student Testing Program in Reading, Writing, and Mathematics, Source: www.doe.state.de.us.*

[xxi] cite DOE Proficiency levels report
[xxii] *High Stakes: Testing for Tracking, Promotion, and Graduation, p. 276*

[xxiii] Fuhrman & Elmore
[xxiv] Fuhrman & Elmore, p.279
[xxv] The Missing Link, p. 50/
[xxvi] Fuhrman & Elmore, p. 276