Definitions for the *No Child Left Behind Act of 2001:* Assessment

Judith Wilde, PhD
National Clearinghouse for English Language Acquisition and
Language Instruction Educational Programs
The George Washington University
Washington, DC

January 2004

The National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs (NCELA) is funded by the U.S. Department of Education's Office of English Language Acquisition, Language Enhancement and Academic Achievement for Limited English Proficient Students (OELA) and is operated under contract No. ED-03-CO-0036 by The George Washington University, School of Education and Human Development. The contents of this publication do not necessarily reflect the views or policies of the Department of Education, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Table of contents

List of exhibits

Introduction

Assessment is a fearful topic to many people, even those involved in education on a regular basis. The purpose of this document is to explore the world of assessment within the context of the *No Child Left Behind Act of 2001*. The definitions provided will help those with little knowledge of assessment to understand the essentials of practice and theory. The information is meant for assessment users so they can interpret the purpose of tests and test scores appropriately and explain them to others. It is not meant to provide in-depth knowledge. The document also provides background for other NCELA documents in the *Definitions for No Child Left Behind* series: *Scientifically-Based Research*, *Research and Evaluation that Work within NCLB Standards*, and *Criteria for Evaluating Evidence-Based Research.*

No Child Left Behind Act of 2001

On December 13, 2001, the 107th Congress passed the *No Child Left Behind Act of 2001* (NCLB), the latest reauthorization of the *Elementary and Secondary Education Act of 1965* (ESEA); President George W. Bush signed the legislation in January 2002. With this legislation, Congress and the President encourage the use of annual assessment of all students to promote high quality education. Both *Title I: Improving the Academic Achievement of the Disadvantaged* and *Title III: Language Instruction for Limited English Proficient and Immigrant Students* include statements about measuring language proficiency and academic achievement using high quality assessments. These mandates represent an opportunity for states and districts to develop and maintain a full assessment system that meets their own needs as well as those of the federal Department of Education. This assessment system must include multiple, up-to-date, quality measures that encompass a reasonable portion of the curriculum. The assessments must be

- valid, reliable, and fair to all students;
- available to students in a manner that allows them to show what they know; and
- available in the student's home language for at least the first year of schooling in the US.

There is no doubt that we must assess all students in order to determine their educational progress. Assessment provides important information for accountability to teachers, administrators, parents, the community, and the students themselves. However, we must be careful to align the need for accountability with quality instruction and assessments, providing a system that is appropriate for all students. When accountability includes high-stakes decisions about grade promotion/retention, placement in core content classes, or academic achievement, it is imperative that the system address the unique characteristics and needs of all students: students living in poverty, English language learner (ELL) students, culturally diverse students, and so on. Only then can we determine the best way to assess all students' achievements and make instructional decisions.

Assessment

Assessment is a broad term that involves the collection and maintenance of various types of data about students including norm-referenced tests, criterion-referenced tests, classroom-based assessments of various types, and performance-based tasks. We use the term "assessment" throughout this document to refer to any situation in which students must respond to items or tasks in order to demonstrate their knowledge and/or skills in a specific area. Using the appropriate type of assessment for a specific purpose is important to the validity and fairness of that assessment. A particular assessment can be reliable, valid, and fair for one purpose, but not for another. For instance, the Iowa Test of Basic Skills (ITBS) may be valid, reliable, and fair for measuring language arts achievement, but not for measuring English language proficiency.

An assessment system must include the technical standards of validity, reliability, and fairness as well as respond to issues of bias and the interpretation and use of assessment results. Different

types of assessments can and should be used within an appropriate assessment system. Each assessment must be considered carefully and should be related to other assessments in order to provide a thorough picture of each individual student. Each assessment should provide the classroom teacher, the school, and/or the school district with information about the students they are serving and, by implication, the teachers who are working with the students. Though it does not specifically mention different types of assessments, the *No Child Left Behind Act of 2001* is clear in referring to multiple assessments (usually by mentioning "assessment**s**" [emphasis added]). Some examples from Title I and Title III of NCLB include:

- assessments that "involve multiple up-to-date measures of student achievement, including measures that assess higher-order thinking skills and understanding" (§1111(3)(C)(vi)),
- ensuring that high quality assessments are used, including those that are valid and reliable (§1001(1), §1111(b)(2)(A)(i), §1111(b)(2)(D)(ii), §1111(b)(3)(C)(iii-iv), §1112(b)(1)(A), §3121(a)(3)),
- using multiple up-to-date assessments (§1111(b)(3)(C)(vi)),
- other academic indicators such as State or locally administered assessments (§1111(b)(2)(C)(vii), §1116(a)(1)(B), §1111(b)(4)),
- assessments in various languages (§1111(b)(6), §1111(b)(3)(C)(ix)(III)),
- assessments of English language proficiency (§1111(b)(7), §3121(a)(3), §3121(d) (1-2)), and
- assessments of various content areas (§1111(b)(3)(A), §1111(b)(3)(C)(v)(I), §1116(b)(3)(A)(ii)).

Academic Achievement

Academic achievement refers to students' concepts, skills, and knowledge in the core content areas of reading and/or language and math as well as science and history or social studies. In order to be successful in academic areas, students must have (1) the opportunity to learn the material and (2) the opportunity to demonstrate that they know the material. Academic achievement is specifically classroom-based, taking place within schools. Several sections of both Title I and Title III of NCLB refer to assessing students' achievement in core content areas; specifically, districts and states are responsible for

- ensuring that core academic subjects are assessed in manners that are appropriate for all students (§1111(b)(3)(C)(ix) and §1116(b)(3)(A)(ii)),
- ensuring valid and reliable assessments of academic achievement standards (§3121(d)(2) and §3121(a)(3)), and
- states "shall make every effort to develop" assessments in other languages (§1111(b)(6)).

As a first layer of definition, a *norm-referenced*, *criterion-referenced*, or *alternative-based* assessment of academic achievement describes current levels of knowledge, attitudes, proficiencies, and skills.

Norm-referenced tests

Standardized assessments can be used to measure participants' skills and knowledge. They are so named because administration, format, content, language, and scoring procedures are the same for all participants – these features have been "standardized." Locally-developed and commercially-available standardized assessments have been created to assess achievement in most content areas. Items generally are multiple-choice or the newer extended multiple-choice which require a student to select the correct response option and then to justify why it is correct. When considering the definition of "standardized test," it is clear that all high-stakes tests should be standardized to some extent, whether they are commercially available or locally developed.

When referring to standardized assessments, most people think of ***norm-referenced tests*** (NRTs). NRTs typically are used to sort people into groups based on their assumed skills in a particular area (for instance, those in the top 10% of skills). They are useful when selecting participants for a particular program because they are designed to differentiate among test-takers. In addition, NRTs can provide general information that will help to match classrooms for overall achievement levels before assigning them to a particular program.

Publishers of nationally developed NRTs administer them to many students across the nation as part of the development process. These students become the ***norm group***. The students are selected to be "typical" of students across the nation who are receiving a "conventional" curriculum and should represent populations with whom the test ultimately will be used. Anyone using NRTs should check with the publisher's technical manual for a description of the norm group since norm groups provide a standard against which the local schools can compare their own students. For instance, if a school's population includes African-Americans, Chinese-Americans, and English-speaking White students, then administrators at the school will want to ensure that these three ethnicities/races were part of the norm group; if they were not in the norm group, then the appropriateness of the NRT for this school's population of students may be questionable. In addition, school staff should ensure that the purposes of the assessment and the content of the assessment match the goals and content of the school's curriculum. Typically, the match between local content and NRT items is less than 50%, but since districts across the country face the problem of aligning cur-

ricula, assessments, and standards, this is improving. When using a nationally available NRT, the technical manual should be reviewed for this and other information.

Criterion-referenced tests

Criterion-referenced tests (CRTs) measure how much or whether specific knowledge has been gained; that knowledge is the criterion against which the participant is measured. CRTs usually contain multiple-choice items, or the newer extended multiple-choice items. Responses generally are marked as "correct" or "incorrect." A score of 80% correct usually is considered as mastery of the knowledge being measured; a score of at least 50% correct indicates that a participant has sufficient understanding of the content to move on to the next level of information or topic to be learned.

CRTs must be aligned closely to the curriculum (which of course must be aligned closely to the district or state content standards) in order to ensure that what is being tested is what has been taught. CRTs should be used before the content is taught, then repeated after the content is taught, thus ensuring that students' knowledge is based on what was taught; i.e., that they did not know the content before instruction.

Attempts have been made to create assessments that could be used in both norm-referenced and criterion-referenced manners. These attempts generally fail because the purposes of the two types of assessment are so different.

Alternative assessments

Alternative assessments are types of measures that fit a contextualized measurement approach, meaning that they can be incorporated easily into classroom routines and learning activities. Their results are indicative of the participant's performance on the skill or subject of interest. As used within this document, "alternative assessment" subsumes authentic assessment, performance-based assessment, informal assessment, ecological assessment, curriculum-based assessment, and other similar forms that actively involve the participant. Some example activities and products that can be used as alternative assessments are listed in Exhibit 1.

Exhibit 1: Example activities that can be used as alternative assessments

- Essays and reports
- Journal entries and logs
- Posters, artistic media
- Brainstorming
- Observations
- Questionnaires
- CRTs
- Poetry and creative writing
- Collaborative work
- Reading lists
- Writing samples
- Anecdotal records
- Cloze tests
- Teacher and student checklists
- Story retelling
- Homework
- Games
- Debates, presentations
- Peer reviews
- Miscue analysis

Alternative assessments often are referred to as measuring whether the student can *think* because they generally do not ask the student to identify a correct answer, but rather to consider the information they have, modify their knowledge, and then apply it to a specific problem. In this way, alternative assessments tap higher order thinking skills more frequently than do multiple choice NRTs.

Note: *Alternative* assessment should not be confused with *alternate* assessment. The former can be used with any population for whom we want to show progress, it is an alternative to formal, standardized testing. The latter is a type of assessment used with populations who cannot complete the content or format of assessments used with mainstream populations.

Language proficiency

Language proficiency definitions vary by state but generally refer to both productive (speaking, writing) and receptive (reading, listening) skills, as first recommended by the Council of Chief State School Officers (1992). Assessing language proficiency is a difficult issue. Language proficiency assessments

- must be appropriate for students of different cultural, ethnic, social, and educational backgrounds;
- are assumed to be able to predict how well a student will do in academic classes although they do not include information about cognitive abilities or academic achievement; and
- tend to measure specific aspects of language (e.g., word choice, grammar) rather than overall communicative competence, which has a repertoire of communication skills that can be used in a variety of situations.

Students can acquire English through multiple sources such as school, playground, church, television and radio, as well as neighborhood children. It is important that language proficiency assessments measure not only a conversational level of English, but also the academic English necessary to function on grade level in all-English-language classrooms.

Both Title I and Title III refer to assessing students' language proficiency. Specifically,

- schools, districts, and states are responsible for ensuring valid and reliable assessments of English proficiency (§3121(a)(3) and (d)(1)) and
- Local Education Agencies must "…provide for an annual assessment of English proficiency (measuring students' oral language, reading, and writing skill in English) of all [ELL] students" (§1111 (b)(7)).

Testing the four modes

There was a time when students' skills in speaking English were the sole measure of their ability to participate in an English-speaking, mainstream classroom. In 1992, the Council of Chief State School Officers formally introduced the concept of assessing all four modes of language proficiency, listening, speaking, reading, and writing, in order to ensure that students' English language capabilities would allow them to participate fully in the English-speaking, mainstream classroom. Some language proficiency instruments measure only one mode (e.g., the Student Oral Language Observation Matrix [SOLOM]) while others measure all four modes (e.g., the Language Assessment Scales [LAS] if using both the LAS-Oral and the LAS-Reading/Writing or the IDEA Proficiency Test [IPT]). Each assessment instrument is somewhat different and must be reviewed carefully to determine whether its cost, time, administration, and purpose meet the needs of the district. More importantly, language arts achievement tests (e.g., the SAT-9 or TerraNova) cannot be used to assess language proficiency; language arts achievement is not language proficiency.

Comprehension

NCLB states that the four modes of English proficiency must be assessed as well as "comprehension." At the present time, it is not clear exactly what is meant by "comprehension." The following definitions have been suggested and should be considered by those selecting language proficiency assessments for students.

1. A student whose average score on the assessment of listening and reading is above a "limited English proficient" level is demonstrating "comprehension."
2. A student who understands what is happening in class, follows directions, and generally seems to grasp the overall pattern of the class, is demonstrating "comprehension."
3. A student whose language proficiency testing indicates that s/he is proficient in English is, by definition, demonstrating "comprehension."
4. A student whose average score on the assessment of all four modes is above a "limited English proficient" level is demonstrating "comprehension."
5. A student who can respond to basic questions (e.g., name, age, description of family members) prior to the administration of a language proficiency test is demonstrating "comprehension."
6. A student who can follow the instructions for an assessment, and understand its purpose and importance, is demonstrating "comprehension."

It does not appear that a specific test of comprehension is required, so schools, districts, and states should work together to define "comprehension" in a manner that meets their expectations of students in English-speaking classrooms.

Scoring mechanisms

Scoring an assessment is nearly as important as ensuring that the assessment is valid, reliable, and fair. NRTs, CRTs, and alternative assessments can be scored in several different ways, but the scores are only as helpful as they are understandable and useful. The interpretation of scores can be confusing and can lead to erroneous conclusions about students' performances. Some of the more often-used scoring mechanisms are defined briefly below.

Scores for NRTs

Raw scores provide the number of items answered correctly. These numbers can be manipulated mathematically to give an average correct score for a classroom or a grade level. An average that includes other grade levels is only possible, however, if all students took exactly the same test; an average across different tests is not appropriate since the tests will have different numbers of items, cover different content areas, and have different levels of difficulty.

Grade equivalents or *grade placement scores* indicate how well a student is doing relative to other students in the same grade. They are stated in tenths of a school year (assuming 10 months in a school year), so 7.3 indicates the third month of seventh grade. These scores are extrapolated; they only estimate the relationship between grade levels and test scores with the assumption that students gain knowledge in a predictable upward fashion. Test publishers generally do not test students with a given form of the assessment during every month of each year of school so, again, we do not know exactly how students score during each month of a given grade level, or how students from different grade levels would fair on the test.

These scores are frequently misinterpreted because they are based on the tenuous assumptions that

(1) what is being tested is studied by students consistently from one year to the next,

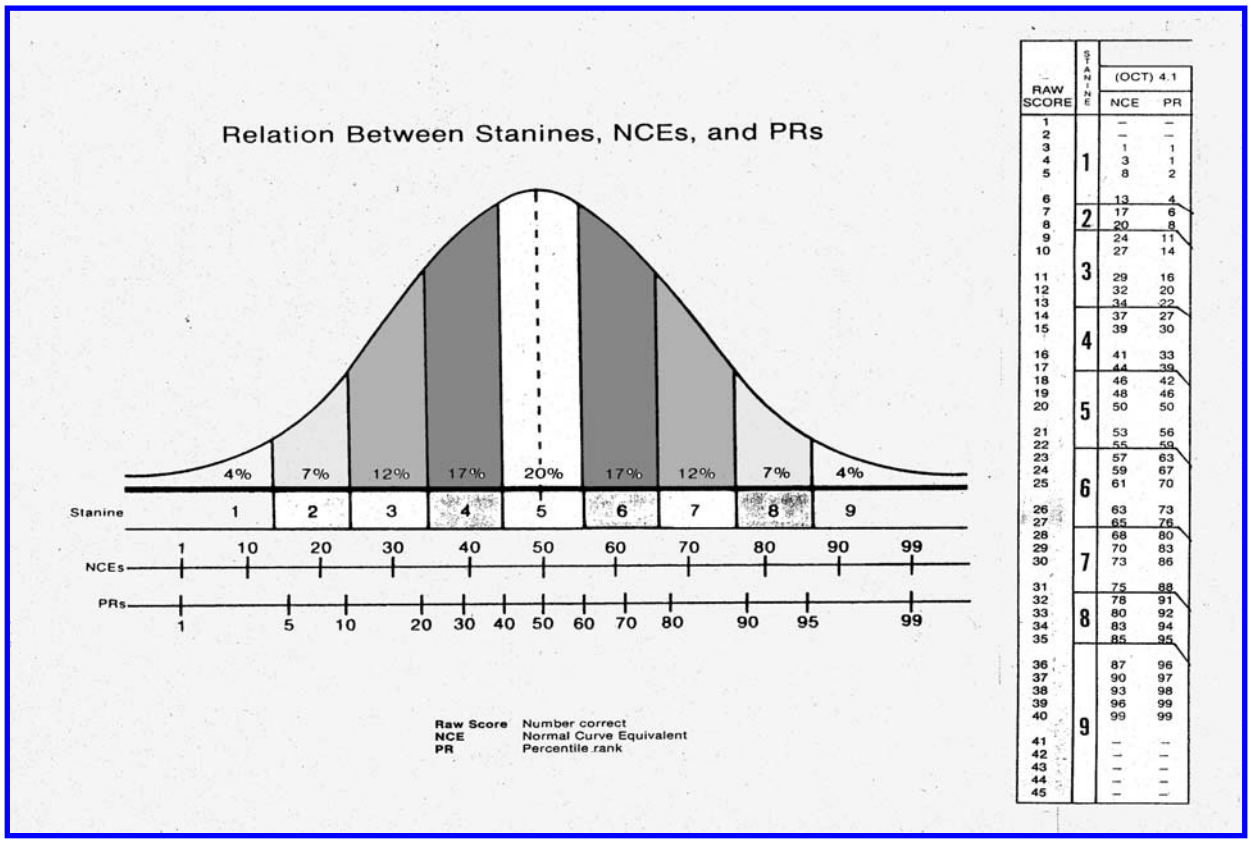
(2) a student's increase in competence is essentially constant across the years, and

(3) tests reasonably sample what is being taught at all of the grade levels for which scores are being reported.

In addition, grade equivalents cannot be summed, averaged, or combined in any way because they are not true numeric scores but represent grade levels – a category rather than a score. They are best used to indicate that students are performing at or near grade-level expectations, above expectations, or below expectations – with no further interpretation or estimate of how much above/below expectations the student scored.

Stanines provide a rough approximation of an individual's performance relative to the performance of other students. Originating from the term "standard nine," stanines divide the range of scores on a test into nine equal groupings. The score of 1 stanine represents the lowest of the nine groups and a 9 represents the highest scoring group. Because of the general nature of stanines, these gross descriptors may communicate individual test results, but using stanines to report group data misrepresents the precision of the data-gathering instruments and forms. As an example, the first, lowest, stanine may include raw scores from 1 to 30. This broad grouping of scores is not only an imprecise method of reporting, but also will make it difficult to show increases – a student may move from a raw score of 2 correct to a raw score of 28 correct, but remain in the lowest stanine. Stanines cannot be used for describing the average achievement level of a class or a group of students; they cannot be averaged, summed, or combined in any way because of the number of raw scores subsumed in each stanine. Exhibit 2 is a graph that indicates the relationship between stanines several scoring mechanisms, including stanines, and the normal curve.

Exhibit 2: Scores on a normal curve



RAW SCORE	STANINE	(OCT) 4.1 NCE	PR
1		–	–
2		–	–
3		1	1
4	1	3	1
5		8	2
6		13	4
7	2	17	6
8		20	8
9		24	11
10		27	14
11	3	29	16
12		32	20
13		34	22
14		37	27
15	4	39	30
16		41	33
17		44	39
18		46	42
19		48	46
20	5	50	50
21		53	56
22		55	59
23		57	63
24		59	67
25	6	61	70
26		63	73
27		65	76
28		68	80
29		70	83
30	7	73	86
31		75	88
32		78	91
33	8	80	92
34		83	94
35		85	95
36		87	96
37		90	97
38		93	98
39		96	99
40	9	99	99
41		–	–
42		–	–
43		–	–
44		–	–
45		–	–

Percentiles are used frequently, and are misinterpreted frequently. They range in value from 1 to 99, indicating the percentage of students scoring at or lower than the test score in question. For example, a student scoring at the 70th percentile scored equal to or better than 70% of the students who took the test; s/he scored higher than 69% of the others. Percentiles are designed to match a normal curve, so 50% of students should score above the 50th percentile and 50% should score below the 50th percentile. Percentiles cannot be averaged, summed, or combined in any way because they are not equal-sized units. As demonstrated in Exhibit 2, percentiles at each end of the normal curve (e.g., scores 1-5) are much larger than those at the middle of the normal curve (e.g., compare scores 1-5 with scores 40-50).

Percentiles should be used only to describe how a student is scoring in relationship to her/his peers. An additional problem with percentiles is that they have no relationship to the actual score achieved. That is, a student may score 29 correct out of 100 items, but if everyone does poorly on the assessment, this low actual score still could put the student at the 90th percentile – looking falsely high.

NRTs often report the percentage of students scoring above the 25th, 50th, and/or 75th percentile. In such cases, it may be possible to show some growth by reporting these numbers. As an example, an evaluation may indicate that "in the 2001-2002 school year, 30% of the 4th graders scored above the 50th percentile but in the 2002-2003 school year, 50% of this same group (now 5th graders) scored above the 50th percentile, indicating a growth in achievement."

There is a whole set of scoring types referred to as ***standard scores***. All standard scores are re-calculated raw scores. As standard scores, each has a predetermined average and a predetermined standard deviation (a measure of how much the scores vary within the group of students taking the assessment – a small standard deviation indicates that the group scored similarly while a large standard deviation indicates that the group's scores were very heterogeneous). A particular type of standard score is the ***normal curve equivalent*** (NCE). NCEs have a national average of 50 and a standard deviation of 21.06; they range in value from 1 to 99 and match percentiles and the normal curve at 1, 50, and 99 (see Exhibit 2). NCEs can be mathematically manipulated (added, averaged, and combined) because the scores are equal-sized units (see Exhibit 2). NCEs do allow careful comparisons across students, across years, across content areas, and, to some extent, across tests from different publishers. An NCE score below 20-25 on a multiple choice test with 4-5 response options for each item, is referred to as a "chance" score; that is, students could achieve this score without looking at the items on the assessment, but just randomly picking one of the response options.

Another frequently used standard score is the ***scaled score***. Various test publishers have created their own unique scales that cannot be described in great detail here. Suffice it to say that these are appropriate scores for use in an evaluation, but care should be taken when comparing the scaled score of one test to the scaled score of another test.

Finally, ***gain scores*** are used to show how much students have progressed since a previous testing period. The usual method for calculating gains is to subtract a previous score from the current score, or a pretest from a posttest. This is problematic because no single assessment is perfectly valid, reliable, and fair. When gain scores are created, all of the technical problems in both the first testing and the second testing are contained in the single gain score, thus making it, in essence, doubly unreliable and potentially invalid and unfair. As an additional problem, the gain score may indicate progress, but does little to indicate how well the student actually is scoring (i.e., what does it mean to indicate that the average student gained 4 points?).

Scores for CRTs

Raw scores can be used to show mastery of an area (e.g., 8 of 10 items answered correctly indicates mastery). A more useful score often is the ***percentage correct***, which provides more information, with a standard understanding of what the score means. Percent correct, when used to measure mastery, can be used within an evaluation to compare students on the same assessment, but cannot be used to compare scores across assessments (e.g., a score on a math achievement test cannot be compared to a score on a language arts achievement test) because the content and difficulty level are different.

Other scoring types that were described in the section on NRT scores can be used with CRTs. However, this would make it difficult to compare a student's score against a specific criterion, which usually is the purpose of a CRT.

Scores for alternative assessments

For many types of alternative assessments, different scoring methods can be used. Three typically used scoring techniques are

- ***Holistic scoring***, which provides a general, overall score for a piece of work – usually with a range from 0-3 up to 0-10 points;
- ***Primary trait scoring***, which defines particular features (or traits) of a performance and then provides separate scores for each trait (e.g., spelling, grammar, voice, sentence structure, and content are some of the traits of writing), usually each scored with values from 0 to 4 or 5; and

- ***Analytic scoring***, which assigns a weight based on the importance of each trait (e.g., sentence structure may be twice as important as spelling).

Exhibit 3 provides example uses for each of these scoring rubrics.

Exhibit 3: Using rubrics

Students have completed a writing sample. The instructor can score the writing sample using holistic, primary trait, or analytic methods, depending on his/her purpose.

Holistic score: the writing sample will be scored on a basis of an overall score of 1-5 (a zero indicates no response at all). Descriptors need to be created for at least the 1-, 3-, and 5-point scores.
1 = limited evidence of achievement, writing skills below grade level
3 = adequate achievement, a general ability to write at an acceptable level for this grade level
5 = excellent achievement, writing is above what would be expected for this grade level
These scores would be applied to the writing sample as a whole.

Primary trait score: the writing sample is reviewed for specific elements (or traits) of writing. Each element is scored separately and, as the class focuses on different elements across the school year, what is scored can be changed. The scores could be the same 1-5 (again with 0 indicating no response at all) described above. The teacher might then score grammar, word choice, spelling, sentence structure, ideas/content, and conventions(e.g., indent each paragraph, capitalize first letter in sentence). A student then would receive a score for each trait, or a total score ranging from 6 to 30.

Analytic score: The scoring of the writing sample builds on primary trait scoring. The same rubric (1-5) can be used, and the same traits can be scored. However in analytic scoring each trait is given a value. For instance, spelling may be considered less important, with a value (or weight) of 1, with grammar of moderate importance with a value of 2, and ideas of great importance with a value of 3. The scoring for a particular child then might be as in the chart below.

Trait	Score	Value/weight	Total score
Spelling	3	1	3
Grammar	2	2	4
Word choice	4	2	8
Sentence structure	4	2	8
Ideas/content	5	3	15
Conventions	4	1	4
Total score for writing sample			42

Scores should be maintained separately for each trait (in order to determine progress in each area) and can be summed to provide an overall score.

Scores for language proficiency assessments

Many language proficiency assessments offer a variety of scoring mechanisms; they can be scored using NCEs, percentiles, raw scores, and so on. Most typically, however, raw scores are converted to a categorical score, or a level of proficiency such as "fluent English speaker," or "competent English writer." There are two problems with using categorical scores:

1. Many raw scores "fit" in each category, making it difficult to see smaller increments of growth; and
2. Using cut-off scores to create categories means that a student's language proficiency can be under- or over-represented based on one item missed or one item guessed correctly.

Raw scores should be collected and maintained for evaluative purposes, but categories can be used to describe the language proficiency of groups of students.

Using and reporting test scores

Many test scores are meaningless when presented without supporting information. As an example, stating that "the average score on a test was 35" or "Joel's score was 40" has little impact. However, adding that "on a test with a possible score of 0-50, these students' scores ranged from 28 to 45 with an average of 35 and Nancy scored 42" gives a great deal more information. This type of information should be provided whenever reporting scores.

An additional problem in interpreting scores may be the type of score presented. However, it is possible to change the type of test score that is reported. That is, if test scores have been recorded in students' files as percentiles, it is possible to change these to more usable scores such as NCEs. Most test manuals will provide a conversion table that includes typically reported scores, such as raw scores, grade equivalents, percentiles, NCEs, stanines, and so on. The table provides equivalencies among the scores. Exhibit 4 provides an example conversion table for scoring mechanisms. By reading across information in Exhibit 4 that describes student performance in the first month of fourth grade (i.e., grade 4.1), scores can be transformed from a raw score of 11 to a stanine of 3, an NCE of 30, and a percentile of 17. The final columns indicate that a score of 11 is a grade equivalent (GE) of 2.7 and an extended scale score (ESS) of 430. Note that because the raw test scores range from 1 to 45, and percentiles and NCEs range from 1 to 99, some percentile and NCE scores are not on the table (e.g., the jump from 8 NCEs to 13 NCEs or from the 33rd percentile to the 39th percentile).

Exhibit 4: Example conversion table for scoring mechanisms

Grade 4								Grade 5								All Grades	
Raw Score	Stanine	October Grade 4.1		February Grade 4.5		May Grade 4.8		Raw Score	Stanine	October Grade 5.1		February Grade 5.4		May Grade 5.8		Raw Score	GE	ESS
		NCE	PR	NCE	PR	NCE	PR			NCE	PR	NCE	PR	NCE	PR			
1	1	–	–	–	–	–	–	1	1	–	–					1	–	342
2		–	–	–	–	–	–	2		–	–					2	–	351
3		1	1	–	–	–	–	3		–	–					3	–	361
4		3	1	1	1	–	–	4		–	–					4	–	370
5		8	2	2	1	1	1	5		1	1					5	–	380
6		13	4	7	2	3	1	6		5	2					6	–	389
7	2	17	6	11	3	7	2	7		9	3					7	–	399
8		20	8	15	5	11	3	8		12	4					8	–	406
9	3	24	11	19	7	15	5	9	2	15	5					9	–	415
10		27	14	22	9	19	7	10		19	7					10	2.5	424
11		30	17	26	13	23	10	11		21	8					11	2.7	430
12		33	21	29	16	26	13	12	3	24	11					12	2.8	437
13	4	36	25	32	20	29	16	13		27	14					13	3.1	443
14		39	30	35	24	32	20	14		29	16					14	3.3	449
15		41	33	37	27	34	22	15		31	18					15	3.5	454
16		44	39	39	30	36	25	16		33	21					16	3.7	459
17	5	46	42	41	33	38	28	17	4	35	24					17	3.8	464
18		48	46	43	37	40	32	18		37	27					18	4.0	469
19		50	50	45	41	42	35	19		39	30					19	4.1	473
20		52	54	47	44	43	37	20		41	33					20	4.2	478
21		54	58	49	48	45	41	21		43	37					21	4.4	482
22	6	56	61	51	52	47	44	22	5	45	41					22	4.5	487
23		58	65	53	56	49	48	23		47	44					23	4.7	491
24		69	68	55	59	51	52	24		48	46					24	4.9	496
25		62	72	57	63	53	56	25		50	50					25	5.1	500
26		64	75	59	67	55	59	26		52	54					26	5.3	504
27	7	66	78	61	70	57	63	27		53	56					27	5.5	508
28		69	82	63	73	60	68	28		55	59					28	5.6	512
29		71	84	66	78	62	72	29	6	57	63					29	5.8	517
30		73	86	68	80	64	75	30		59	67					30	6.0	522
31	8	75	88	70	83	66	78	31		61	70					31	6.2	527
32		78	91	72	85	68	80	32		63	73					32	6.5	533
33		80	92	75	88	71	84	33	7	66	78					33	6.7	538
34		83	94	77	90	73	86	34		68	80					34	7.0	545
35		85	95	80	92	76	89	35		71	84					35	7.3	551
36	9	87	96	82	94	79	92	36		73	86					36	7.5	557
37		90	97	85	95	82	94	37		76	89					37	7.8	564
38		93	98	88	96	84	95	38	8	80	92					38	8.3	573
39		96	99	91	97	88	96	39		83	94					39	8.7	581
40		99	99	95	98	91	97	40	9	87	96					40	9.2	590
41		–	–	99	99	96	99	41		91	97					41	9.8	600
42		–	–	–	–	99	99	42		95	98					42	10.4	611
43		–	–	–	–	–	–	43		99	99					43	11.1	623
44		–	–	–	–	–	–	44		–	–					44	11.9	637
45		–	–	–	–	–	–	45		–	–					45	12.8	652

... AND SO ON ...

It is clear from these brief descriptions that even some of the more common scoring techniques are not particularly useful for evaluation purposes. Exhibit 5 lists the scoring types discussed above and indicates whether or not they can be used to describe general performance and/or can be used in computations for an evaluation. In several cases, comparisons can be made to the norm group, the large number of students who took the assessment at the behest of the test publisher. These students form the "standard," it is their scores that establish percentiles, NCEs, and so on. If a program is hoping that their students will "look like," or have scores similar to, the national average score on a particular NRT, then the norm group is an appropriate comparison.

Exhibit 5: Test scores and their uses

Type of score	Score compares students against	Scores can be used for	
		Evaluation	Description
Raw scores	Nothing, there is no comparison	Yes	Yes
Percent correct	Standard of 100% correct	Yes	Yes
Grade equivalents	Norm group	No	Perhaps
Standard scores, including NCEs	Norm group	Yes	Yes
Stanines	Norm group	Not suggested for any purpose	
Percentiles	Norm group	No	Yes
Gain scores	Their own previous score	No	Perhaps
Mastery scores	Criterion for acceptable skills, knowledge	Yes	Yes
Categories or levels	Norm group considered in each level	No	Yes

Finally, a comment on cut-off scores. Cut-off scores typically are used to categorize scores into groupings of similar students (e.g., high, medium, and low readers or students who have attained below basic proficiency, basic proficiency, proficiency, or advanced proficiency levels). These scores can be harmful to the life of the student when rigidly enforced and used to make life decisions. For instance, consider the recent growth in testing to ensure that students have met minimum criteria to graduate from high school. In most cases, there is one test that is administered to students that is the final determination of graduation status – those who pass, graduate, those who do not pass may have the opportunity to take the test again or they do not officially graduate. We would suggest the following procedures when creating cut-off scores:

1. Consider carefully whether a cut-off score truly is needed;
2. Spend time with consultants, researchers, and staff to determine a cut-off score that is appropriate;
3. Create a "gray area" for borderline students – e.g., if 60% correct is the cut-off score, then scores between 58% correct and 62% correct are in the "gray area;"
4. Determine further assessments or reviews of past accomplishments to determine whether students in the "gray area" pass or fail the assessment; and
5. Communicate these guidelines clearly to students so that they understand the options.

Technical qualities of assessments

Meaningful assessment is essential. To ensure that an assessment is meaningful, three factors must be considered: reliability, validity, and fairness. While psychometricians still argue about the relative importance of each of these concepts and what constitutes "good" reliability, validity, and fairness, some general explanatory statements can help to clarify these assessment qualities. There also are several references within Title I and Title III of NCLB to assessments that are

- "high-quality" (§1001(1), §1111(b)(3)(A), and §1112(b)(1)(A)),
- aligned to state content and achievement standards (§1111(b)(2)(A)),
- improved (§3115(c)(2)(A)), and
- valid and reliable (§1111(b)(2)(D)(1), §1111(b)(3)(C)(ix)(III), and §3121(a)(3)(B)).

Reliability

Reliability is the stability or consistency of an assessment. For instance, two assessments of a student, performed at the same time, should show similar results; two reviews of a teacher's qualifications should result in similar conclusions. An instrument must be reliable if it is to be used to make decisions about how well a participant is performing or how well a staff development program is succeeding. As a general rule, the more items on an assessment, the greater the reliability. An assessment with 50 items will be more reliable than an assessment with 10 items; however, an assessment with 300 items may fatigue the test-takers and be very unreliable. Most psychometricians agree that at least 10 items are needed for each area tested (i.e., the various subareas in a language arts achievement assessment should each have at least 10 unique and separate items) in order to have a reliable instrument.

Reliability is measured on a scale from 0.0 to 1.0, with higher numbers being better (i.e., more reliable) – although it is virtually impossible to achieve a rating of 0 or 1. Most psychometricians agree that a reliability coefficient of at least

- .80 is needed if a test will be used to make decisions about a single individual;
- .65 is needed if a test will be used to make decisions about a group of individuals such as a classroom; and
- .50 is needed if a test will be used to provide some general information about how well a group of individuals is performing.

Validity

Validity is more difficult to describe, in part because psychometricians are changing their own views of validity. The newer view is that validity asks whether the interpretation, uses, and actions based on assessment results are appropriate. It is especially important to consider the communicative competence of learners when creating a valid test. In addition, the specific purpose of the assessment must be considered. An assessment may be valid for one purpose, but not for another. Basic questions when considering validity are "Does this test measure what it purports to measure?", "Do I believe what this test tells me about my learners?", and "Are the results of this assessment similar to results from other assessments of the same topic?"

Fairness

Fairness refers to testing that considers the language, gender, culture, and overall abilities of the test-takers. For instance, if it is known that a group of test-takers have difficulties writing in English, then a fair test will include response options that allow the students to create pictures or graphs to show their answers or that allow them to dictate answers to a fluent English person. Fairness is impacted by how items are developed, the scoring procedures used (as well as the training of scorers and the calibration of scores), access to good instruction, and so on.

Fairness also should ensure that biases are not evident in the testing procedures or test items. Biases generally fall into three areas:

- biases in item development or scoring procedures that unjustly promote or oppose an individual's race/ethnicity, culture, language, or beliefs;
- stereotyping within items or reading passages based on race or ethnicity, language, culture, or physical ability through under/over representing or ridiculing certain groups; and
- illustrations that negate the impact of certain individuals, typically by not including them.

Biases can be quite subtle. For instance, if items on an assessment only use names that are typically associated with Anglos, there is a relatively subtle bias for one group (Anglos) and against others (e.g., Asians, Hispanics) who choose to maintain culturally-appropriate names. These issues can impact students' interest in subject matter as well as their interest in achieving on an assessment.

Conclusions: Making assessment meaningful and useful

There is no doubt that we must assess all students in order to determine their educational progress. Assessment provides important information for accountability to teachers, administrators, parents, the community, and the students themselves. However, we must be careful to align the need for accountability with quality instruction and an assessment system that is appropriate for all students. When accountability includes high-stakes decisions about grade promotion/retention, placement in core content classes, or academic achievement, it is imperative that the system address the unique characteristics and needs of all students: students living in poverty, English language learner students, culturally diverse students, high achievers, and so on. Only then can we determine the best way to assess all students' achievements and make instructional decisions.

Assessment leads to accountability by informing various stakeholders about student progress. In order to ensure that accountability is meaningful, an assessment system should include NRTs, CRTs, and alternative assessments. In addition, there must be a systematic process of identification, placement, continuously monitoring progress, transition from ESL or dual language support into English-only classrooms (if appropriate), and inclusion of all students in the full assessment system. Thus we must align the need for accountability and quality instruction with an assessment system that is appropriate for all students.

NCLB states that scores of assessments of both academic subjects and language proficiency shall be used by the Local and State education agency "for improvement of programs and activities; to determine the effectiveness of programs and activities in assisting [ELL] students to attain English proficiency and meet challenging State academic content and student academic achievement standards…" (§3121(b)(1)(2)). Thus we must always ensure that

- the assessments used with all students are of the best and highest quality possible;
- multiple assessments are used, especially when making major life decisions;
- scores are maintained across time so that progress can be followed carefully; and
- interpretations of scores are made with wisdom and understanding.

In order to follow these mandates, we suggest that the following elements are essential within any assessment system used for accountability purposes:

- use multiple assessments of different types (e.g., an NRT, a CRT, and an alternative assessment);
- ensure that all assessments are reliable, valid, and fair;
- create a policy that indicates when students should be tested in what language(s);
- assure that staff development activities, curricula, expected teaching techniques, and assessments are aligned;
- do not use one assessment, of any type, to make a life-decision (e.g., program placement, graduation);
- provide annual training sessions for those who administer and score assessments;
- maintain long-term data for each student (that is, keep scores from past years as well as this year's scores);
- when using an NRT, read the technical manual to determine the ethno-linguistic groups who participated in the norming process;
- review each assessment – items, paragraphs to be read, response options, and scoring techniques – for biases and stereotyping;

- maintain data in the raw form, or at least in as detailed a form as possible (i.e., do not keep categorical information – it is easy to create categories from the raw data, but not *vice versa;* and
- review assessments often to ensure that they continue to meet the needs and policies of the local school district.

It may be helpful to create a team or advisory panel that helps make assessment decisions. Such a panel should include administrators, teachers, paraprofessionals, parents, and community members; at upper grade levels, students may be included as well. This group could be given the mandate to review

- the alignment of curricula, instruction, and assessment;
- selection and/or development of assessment(s);
- new assessments for bias;
- cut-off scores for specific purposes; and
- generally ensure that assessments are used in an appropriate manner.

This document has been fairly short while attempting to encompass a difficult and complex topic. For those who wish to learn more on any of these aspects of assessment, a reference list and bibliography follows. Some of these books are new, some are "classics" in the field. In addition, most colleges and universities offer classes in "tests and measures" or "psychometrics" in schools of education and/or psychology.

Reference list and Bibliography

American Educational Research Association (July 2000). AERA position statement concerning high-stakes testing in preK-12 education. Washington, DC: Author.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards of educational and psychological testing*. Washington, DC: Joint Committee of AERA, APA, and NCME.

Berman, P., S. Aburto, B. Nelson, C. Minicucci, and G. Burkart (2000). *Going Schoolwide: A resource guide for comprehensive school reform inclusive of limited English proficient students.* Oakland, CA: Institute for Policy Analysis and Research.

California Department of Education (1999). *Designing a standards-based accountability system for language minority and immigrant student populations, 2nd edition.* Sacramento, CA: Author.

Castellon-Wellington, M. (June 2000). The impact of preference for accommodations: The performance of English language learners on large-scale academic achievement tests. CSE Technical Report 524. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Commission on Behavioral and Social Sciences and Education (1998). *High stakes: Testing for tracking, promotion, and graduation.* National Academy Press. Available on-line at www.nap.edu/books/0309062802/html

Council of Chief State School Officers (1992). *Recommendations for improving the assessment and monitoring of students with limited English proficiency*. Alexandria, VA: Author.

Cummins, J. (1981). Language proficiency and academic achievement. In J.W. Oller (ed), *Issues in language testing research*. Rowley, MA: Newbury House.

Cummins, J. (1992). Language proficiency, bilingualism, and academic achievement. In P.A. Richard-Amato & M.A. Snow (eds), *The Multicultural Classroom: Readings for Content Area Teachers.* London: Longman.

Del Vecchio, A., M. Guerrero, C. Gustke, P. Martínez, C. Navarrete, C. Nelson, & J. Wilde (Summer 1994). *Whole-school bilingual education programs: Approaches for sound assessment*. Program Information Guide Series #18. Washington, DC: National Clearinghouse for Bilingual Education.

Durán, R.P. (1988). Validity and language skills assessment: Non-English background students. In H. Wainer and H.I. Braun (eds) *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

FairTest (1995). *Bilingual assessment fact sheet.* Cambridge, MA: Author.

Figueroa, R.A. (1990). Best practices in the assessment of bilingual children. In A. Thomas and J. Grimes (eds.) *Best Practices in School Psychology-II.* Washington, DC: National Association of School Psychologists.

Figueroa, R.A. & S. Hernandez (May 2000). *Testing Hispanic students in the United States: Technical and policy issues.* Washington, DC: President's Advisory Commission on Educational Excellence for Hispanic Americans.

Geisinger, K.F. & J.F. Carlson (July 1992). *Assessing language-minority students.* Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Gottlieb, M. (2000). Standards-based, large-scale assessment of ESOL students. In M.A. Snow, *Implementing the ESL Standards for Pre-K-12 Students through Teacher Education.* Alexandria, VA: Teachers of English to Speakers of Other Languages (TESOL).

HR1513, 107th Congress (April 4, 2001). A bill to provide for fairness and accuracy in high stakes educational decisions for students.

Hamayan, E. (1997). Critical issues affecting the educational success of language minority students: Focus on assessment. Des Plaines, IL: IL Resource Center.

Heubert, J.P. (2000). High-stakes testing: Opportunities and risks for students of color, English-language learners, and students with disabilities. Washington, DC: National Center on Accessing the General Curriculum. Available on-line at www.cast.org/ncac

International Reading Association (July 1999). High-stakes assessments in reading: A position statement of the IRA. Newark, DE: Author.

Joint Committee on Testing Practices (1988). *Code of fair testing practices in education.* Washington, DC: American Psychological Association.

Linn, R.L. (April 2001). The design and evaluation of educational assessment and accountability systems. CSE Technical Report 539. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Menken, K. (September 2000). *What are the critical issues in wide-scale assessment of English language learners?* Issue & Brief #6. Washington, DC: National Clearinghouse for Bilingual Education.

McKeon, D. (November 2000). High stakes testing and English language learners. Washington, DC: National Education Association.

Minnesota Assessment Project (1996). *Accommodations for students with limited English proficiency: Analysis of guidelines from states with graduation exams.* State Assessment Series, Minnesota Report 6. Minneapolis, MN: National Center on Educational Outcomes.

National Academy Press (1998). *High Stakes: Testing for tracking, promotion, and graduation.* Available on-line at www.nap.edu/books/0309062802/html

National Council of Teachers of English (November 22, 2000). English teachers pass resolutions on high stakes testing and the rights of test takers. Urbana, IL: Public Affairs Office, IRA.

National Education Association (NEA) (2001). High-stakes – Testing plus: Real accountability with real results. Washington, DC: Author.

National Evaluation Systems (1991). *Bias Issues in Test Development.* Amherst, MA: Author.

Office for Civil Rights (December 2000). *The use of tests as part of high-stakes decision-making for students: A resource guide for educators and policy-makers.* Washington, DC: US Department of Education.

Popham, W.J. (November 2000). Educational mismeasurement: How high-stakes testing can harm our children (and what we might do about it). Washington, DC: National Education Association.

Rivera, C., C.W. Stansfield, L. Scialdone, & M. Sharkey (April 2000). An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998-1999. Final report. Sponsored by the Office of Bilingual Education and Minority Languages Affairs, US Department of Education. Arlington, VA: Center for Equity and Excellence in Education, The George Washington University.

Rivera, C. & C.W. Stansfield (April 12, 2001). The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

School District of Philadelphia (2001). *Assessment accommodations booklet, SAT-9/Aprenda, PSSA.* Philadelphia: Office of Assessment.

Wiggins, G. (date unknown, about 1993). *An assessment glossary*.

Wisconsin Department of Public Instruction (January 1999). *DPI guidelines to facilitate the participation of students with special needs in state assessments.* Madison, WI: author.

Zehler, A.M., P.J. Hopstock, H.L. Fleischman, & C. Greniuk (1994). *An examination of assessment of limited English proficient students*. Arlington, VA: SIAC. Available on-line at www.ncbe.gwu.edu/miscpubs/siac/lepassess.htm