Redmond
F-6

Schools

What to Look for in the Technical Manual: Twenty Questions for Users

Thomas A. Hutchinson
Applied Symbolix, Chicago, IL

Most standardized tests published for speech-language pathologists include details concerning the technical quality of the instruments. Typically, this technical information includes information concerning the sample to whom the test was administered, the norming procedures, and the reliability and validity of the test. Most users expect to see this information somewhere in a test manual, but it is doubtful that very many give the same amount of attention to these technicalities as they give to the mechanics of administration directions and scoring rules.

A prospective buyer of a new car is more likely to kick the tires or sit behind the wheel than to look under the hood. A prospective user of a new or unfamiliar test is more likely to evaluate the test by surveying its content than by reviewing its technical details. A systematic content survey can reveal much about a test if it answers the following questions:

- How long does the author claim it will take to administer and score the test?
- What does the author give as the purpose and uses of the test?
- Are the names and descriptions of the subtests recognizable and do they make sense?
- Are the items, directions, and scoring rules clear and practical?
- Are the norms reported in terms (e.g., percentiles, standard scores, age equivalents) that match the local regulations and current practices?

ABSTRACT: The details presented in technical manuals for tests generally address a common core of issues related to the psychometric quality of the tests and the interpretation of their results. In this article, major categories of technical information are described in practical terms and related to test use and interpretation: (a) logical evidence of validity, (b) empirical evidence of validity, (c) types of reliability estimates for evaluating and interpreting tests, and (d) practical issues in understanding standardization data and using norms. The article is organized as a series of 20 questions concerning these categories of information. Responses to the questions include discussions of key measurement concepts and examples of the kinds of tests used by speech-language pathologists. Potentially unfamiliar or confusing terms are italicized when first used, and each term is followed by a brief definition.

KEY WORDS: technical manual, validity, reliability, test interpretation, assessment

Although this kind of quick survey can provide most busy professionals with an informative match between a new test's content and their current needs, a similar survey of the technical section offers no such useful information. For many practitioners, then, the tables of numbers and paragraphs of psychometric jargon remain unexamined under the hood.

The purpose of this article is to provide professionals with an aid in navigating the technical sections of test manuals. The article is organized in two levels: (a) a series of topics similar to those addressed by most technical sections or manuals; and (b) under each topic, a series of questions concerning tests, with accompanying reminders of key measurement concepts and practical definitions when appropriate. The topic headings parallel the major headings or chapter titles in technical sections or manuals (see the model Table of Contents in Table 1). However, the order of the sections in this article differ from the order of topics in most manuals and the sections are presented instead in the order of their importance: validity, reliability, standardization, and norming. Potentially unfamiliar or confusing jargon words are italicized when first used in this article, and many terms are followed by brief definitions.

Using this article as guide, a person should be able to turn to the corresponding heading in the technical manual for almost any unfamiliar test and determine whether it addresses the key questions posed here. If a person knows what to look for, where to look, and why it is important, he or she can learn a great deal about a test in as little as half an hour. More important, one can also find much meaning in the numbers—meaning that can contribute as

Table 1. Typical contents of a technical section or manual.

Theory and rationale
Need
Intended uses
Model
Development of the test
Pilot
Tryout
Standardization and norming
The standardization sample
Standardization procedures
Development of norms
Reliability
Internal consistency
Test–retest reliability
Interrater reliability
Standard error of measurement
Validity
Content validity
Criterion-related validity
Concurrent validity
Predictive validity
Construct validity
Intercorrelations
Factor analysis

much to assessing the quality of a test as can knowing its content.

QUESTIONS CONCERNING THE VALIDITY OF THE TEST

In simple terms, *validity* is evidence that a test measures what it is assumed to measure. The quality of this evidence has important implications for both the test giver and the test taker because it addresses the fundamental question of a test's worth. Messick (1989) explained that "validity is a matter of degree, not all or none.... Validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 13).

Consider the practical implications of these two principles. First, validity is not some property either present or lacking in a test—it is a matter of the quality and extent of available evidence. When a test is first published, and usually for many years thereafter, its technical section provides virtually the only summary of available evidence to support that it measures what it claims to measure. Second, validity evidence is less a matter of the test than of the *uses* of the test results. Because the uses may vary across settings or individuals tested, an instrument may be more valid for one purpose (e.g., a screening to identify where a child stands among age peers on some task) than for another purpose (e.g., using the same test to determine what intervention or instruction would improve the child's performance). Third, establishing that a test is valid for certain purposes may require evidence derived from logical analysis or experimental data, or both.

One obstacle to comprehending technical sections is the abundance of technical jargon they contain. For example, those sections that treat validity typically make fine distinctions between different types of validity (e.g., content, concurrent, predictive, construct)—see Table 2 for brief definitions of these types. However, more current views of validity regard these "types" as distinctions between different *sources* of evidence rather than between kinds of validity. Thus, each can contribute to understanding how the scores on a test relate to the judgments to be made (Messick, 1989). The discussion of validity that follows organizes validity evidence into two broad types: logical and empirical.

QUESTIONS CONCERNING LOGICAL EVIDENCE OF VALIDITY

1. Is the purpose of the test explicitly stated?

The purpose (or purposes) of the test should be stated simply and explicitly by the author, in the introduction or the technical section or both. The statement of purpose is important because it helps to define the boundaries the test maker has placed around the *construct* of the test—its conceptual, theoretical, or operational definition of what is measured. Prospective users also have a right to both a sound rationale and some concrete evidence of systematic evaluations that support the proposed uses.

2. Is the construct or model explicitly defined, and does it relate to the stated purpose?

To be interpretable, a test should be more than an arbitrary collection of tasks or items. The extent to which those tasks or items relate—not only to each other but also to some common trait—should be predictable from an understanding of its construct. That construct may be expressed as a concept, a theory, or a well-established relationship between the scores on the test and what the test measures. For example, if a test is made up of a

Table 2. Sources of evidence of test validity.

Content	Evidence that items are representative of the content domain(s) sampled by the test or subtest and are relevant to its intended model and purpose.
Criterion-related	Evidence that a test's results accurately estimate the subject's performance on an already accepted measure of the domain (the criterion); includes evidence that the test estimates the subject's present (concurrent) or future (predictive) performance.
Construct	Once used to refer primarily to statistical data supporting the construct of the test; more recent views treat construct validity as including any qualitative or quantitative information (including content and criterion-related information) that supports the test maker's theory or model underlying the test.

battery of subtests, where does the author explain how the subtests relate to each other and to the overall construct? What claims are made for the usefulness of the subtest scores and any other scores made from composites of subtest scores? This information should make a logical case for the test and provide a basis for evaluating its validity.

3. Is there a clear, supportable rationale for the selection of test content?

This question should be answerable from the technical section or from a separate chapter on the test's construction. One reason many test users turn first to the test items and directions for information concerning a test is that the items and directions indicate what is called *face validity*. Face validity is the judgment of validity based on the appearance of *content validity* rather than on a detailed analysis of content. Unfortunately, a test with many good items may look good but be a poor measure because it samples an inadequate range of performance of the targeted ability or skill. Or it may assume a certain cultural background, degree of world knowledge, or proficiency in a language the child does not have, thereby testing content not relevant to the inferences we want to be made about the child. Instead of relying on face validity, test users need a more systematic evaluation of content. In an article in this forum, Sabers provides an example of an in-depth analysis of content validity of tests appearing to measure sentence production.

A useful set of guidelines for assessing content validity was suggested by Kretschmer and Kretschmer (1978, p. 145): "Are the justifications for the definitions and selection of test items clear enough so that a user could generate additional items or exercises to fit the test model?" In short, if one cannot create additional items and scoring rules that fit the author's model, it is difficult to understand how anyone but the test maker could accurately interpret the child's performance or the relationships among the results on different parts of the test. Difficulty with such a task may point to more complex problems of content, however, and even indicate a lack of clarity or integrity in the test maker's construct or in the underlying assumptions about what the test measures.

QUESTIONS CONCERNING EMPIRICAL EVIDENCE OF VALIDITY

Evidence of test validity should go beyond clear rationales, logical specifications for subtests and items, and sensible interpretation schemes. It should also include empirical data that consistently support the test's construct and stated purposes. Messick (1989) stated:

> Almost any kind of information about a test can contribute to an understanding of its construct validity, but the contribution becomes stronger if the degree of fit of the information with the theoretical rationale underlying score interpretation is explicitly evaluated. (p. 17)

As noted above, the term *construct validity* is often associated with statistics that support a test model, but all of the logical and empirical evidence can be brought together under this single concept.

Empirical evidence frequently takes the form of quantifiable relationships among scores on different parts of the test (called *intercorrelations*) or scores on the test and other tests (called *correlations*). The most common device for reporting these relationships is a *correlation coefficient*, which quantifies such relationships based on the scores of a particular group of subjects. Test users who wish to make sense of the empirical evidence for validity (or reliability) need to know how the correlation coefficient describes this relationship. For this reason, Table 3 provides an overview of the correlation coefficient and the level of significance.

4. What evidence is given to describe the relationship between this test and others considered to be similar?

If a test is developed to identify students with phonological delays, we might want to see evidence that the results agree with those of accepted measures of phonology. This type of *criterion-related validity* evidence uses an existing test as a criterion. A sample of children may be given both tests, with the order of administration being randomly assigned within the group to reduce the chance of some systematic effect of practice with the first test influencing performance on the second test. If the results agree with the results of tests that users already accept as valid measures of the construct, we can have increased confidence in the test. On the other hand, when the test under review produces results that are quite different from the accepted test, some analysis and explanation of the differences in the two tests should be provided.

5. What evidence is given to support the accuracy of this test in classifying subjects into already established performance categories?

To see if the test results agree with the classifications, we might also administer the same phonology test to a sample of children with phonological disorders and an otherwise matched sample of children with normal phonological development. High accuracy in predicting membership in these two groups would support the use of the test for screening or identifying children whose level of phonological development is not yet known. If the test is meant to do more than identify—to indicate specific phonological processes, for example—evidence beyond simple classification into one or the other group should be provided.

6. What statistical data support the relationships among separate components of the test or their relationships with the overall construct?

Users should expect to see some evidence of predictable (or at least interpretable) quantitative relationships between test scores and some other variables they know to be

Table 3. Correlation coefficients and levels of significance.

Name	Correlation coefficient (or coefficient of correlation)
Definition	An expression of the relationship between two numeric values (e.g., scores) obtained from the same group of subjects.
Form	Italic *r* and decimal to one or more places, as in $r = .75$
Values	+1.00 is the highest value, indicating that each person's position in the group is the same for both scores. -1.00 is the lowest value, indicating that each person's position (e.g., highest) in the group on one score has the opposite relationship to the person's position (e.g., lowest) on the other score. 0.00 indicates that there is no relationship between the individuals' positions in the group on the two scores.
Cautions	Note that correlation is *not* causation. Because two variables have a relationship does not mean that one causes the other. Note that a correlation coefficient does not express the proportion or percentage of variance of the two values being correlated. See squared correlation below.

Name	Squared correlation
Definition	The value of a correlation coefficient multiplied by itself. This statistic expresses the proportion of the variance of one variable that can be predicted by the other variable in the correlation.
Symbol	r^2
Values	0.0 to 1.0

Name	Level of significance (or significance level)
Definition	The level of probability that a correlation is 0.
Symbol	Italic *p* and decimal to one or more places, as in $p = .05$
Values	.05 and .01 (meaning that the probability of a correlation being 0 is no greater than 5 in 100 or 1 in 100, respectively) are generally accepted in psychology.
Cautions	Note that the size of the sample is a major variable in determining the significance level. For example, a correlation of .50 is significant at the .05 level if the group has 11 members and significant at the .01 level if the group has 15 members. Even small coefficients may be significant if the group is large enough. For example, if $r = .20$ for a group of 80, $p = .05$.

related to the construct. For example, phonological development is known to change with age in young children. Therefore, one might expect some evidence that test scores concerning a set of stimulus words improve with age, such as those in the fictional data for Group 1 in Table 4. In other words, one looks for statistical evidence that converges with the assumptions of the test maker's model or the inferences that can be logically drawn from the model.

7. What statistical support is presented to describe the relationship between test scores and other scores unrelated to what is measured?

To continue with the example of phonological development, accuracy is not expected to improve much after a certain age (around age 8). In fact, one might question a test that showed steadily increasing scores beyond that age. With students above age 8, a test of phonological development would be expected to yield a pattern of highly similar scores (like those for Group 2 in Table 4), not a pattern of increasing scores (like those for Group 3 in Table 4). Campbell and Fiske (1959) referred to such a pattern of similar and dissimilar results as *convergent–discriminant validation.*

In another example, consider a test of cognitive–linguistic performance designed to assess recovery from traumatic brain injury. If an examiner first administered the test to a child when the injury was acute and then retested the child a few weeks later, notably higher retest scores would be expected. However, no such improvement would be expected from a comparable retest of non-injured peers. The extent to which one can accurately predict and interpret data such as these is important in investigating construct validity.

When a test battery has several subtests, some construct validity support should be evident in reports on the relationships among the various components or subtests. The pattern of relationships also should be at least somewhat predictable from the various elements in the test maker's theory. For example, Table 5 reports the intercorrelation coefficients among four subtests in a fictitious language battery. Judging by the subtest names, the test maker's model of language appears to make distinctions between receptive and expressive modes and between vocabulary and syntax. Table 5 also reports a variable not

Table 4. Average scores on a 50-item test of phonology for three groups of students.

Group 1		*Group 2*		*Group 3*	
Age	*Score*	*Age*	*Score*	*Age*	*Score*
3	22	9	44	9	45
4	29	10	44	10	46
5	35	11	45	11	47
6	37	12	45	12	48
7	39	13	45	13	49
8	43	14	45	14	50

Table 5. Scores of subjects evenly distributed across ages 3–5.

	VC	*VE*	*SC*	*SE*	*HI*
Vocabulary comprehension (VC)	1.00				
Vocabulary expression (VE)	.75	1.00			
Syntax comprehension (SC)	.54	.49	1.00		
Syntax expression (SE)	.52	.57	.69	1.00	
Height in inches (HI)	.25	.33	.29	.34	1.00

$n = 300$. $p < .001$

measured by the test—the child's height in inches (to illustrate another point somewhat later.) As a reminder, the correlation of each variable with itself, which is equal to 1.0, is reported "on the diagonal." Examine the coefficients in Table 5.

8. Are there any statistical patterns that might help a user to better understand the test or better interpret performance?

Table 6 presents the intercorrelations from Table 5, rearranged into three categories identified as *Like content–Different mode*, *Different content–Like mode*, and *Different content–Different mode*. This display, which makes it easier to compare the coefficients across the three categories, indicates that the highest correlations are between subtests with like content (the top row). The lowest correlations are between subtests with different content and different mode (the bottom row). Classifying and rearranging these values makes the pattern clearer. It may also provide some insight for interpreting scores, because it suggests that comprehension and expression appear to be more closely related than are vocabulary and syntax. Even so, the moderate correlation between receptive and expressive indicates that these modes are certainly not independent of each other.

Keep in mind that correlational data do not imply cause–effect relationships, nor do they explain *why* two variables might be related. For example, the correlations between height and language scores in Table 5 are all positive and significant. Does this mean that increasing height contributes to improved language performance, or that developing language causes one to grow taller? Not at all. First, the *significance* of these relatively small coefficients is due to the large number of children sampled (see the caution concerning significance in Table 2). Second, the *magnitude* of these correlations is probably best explained by the fact that both height and language performance tend to increase as children at these ages mature. In fact, at some point (say around age 15), when children have achieved most of their eventual adult height and have developed much of their general language ability, the variability in their height no longer has much relationship with the variability in their language performance.

When evaluating correlations and other expressions of relationships among variables, much of the support for a test's validity comes not simply from the strength of the coefficients but from the strength of the test maker's theory

Table 6. An alternative display of correlation coefficients in Table 5 based on a classification of variables by content and mode.

Like content–Different mode

	Vocabulary expression		Syntax expression
Vocabulary comprehension	.75	Syntax comprehension	.69

Different content–Like mode

	Vocabulary comprehension		Vocabulary expression
Syntax comprehension	.54	Syntax expression	.57

Different content–Different mode

	Vocabulary comprehension		Vocabulary expression
Syntax expression	.52	Syntax comprehension	.49

in predicting the pattern of correlations. When the observed relationships cannot be predicted from the test maker's theory (or other accepted theories concerning the content measured), the validity of the test or the construct should be questioned. In fact, identifying what is *not* related to a construct may also help one better understand it. In the example of height and language performance, the variables are related to each other largely because both increase as children mature. Moreover, we would expect the correlations between these two variables to decrease as children reached adolescence.

Often, a test battery contains so many subtests that a correlation matrix like the one in Table 5 becomes too large to interpret easily. In these cases, a procedure called *factor analysis* is often used to simplify the process of interpretation. Factor analysis is a method of identifying patterns of common *variance* in a matrix of intercorrelations. (The variance of a set of scores is an expression of deviations of scores above and below the mean.) Lack of familiarity with factor analysis may tempt many test users to skip those discussions and tables in a technical manual. However, factor analysis results actually simplify the interpretation of the complex relationships reflected in an intercorrelation matrix, such as the one shown in Table 7.

Although there are different types of factor analysis, results in test manuals are typically reported in a simple format like the one shown in Table 8, which is the result of a factor analysis of the correlation matrix in Table 7. These results present a *rotated solution* made up of rows of variables and columns of *factor loadings*. Usually only one solution is reported, probably the most easily interpreted of several solutions that may have been produced. The names of the variables in the intercorrelation matrix are listed in the left column, and the columns to the right are often headed by Roman numerals ranging from I to the total number of factors.

Table 7. Intercorrelations (decimals omitted) of raw scores of students in Grades 4–6 on a test of written language (n = 885).

	Purpose	*Audience*	*Vocabulary*	*Style*	*Develop*	*Organization*	*Sentence*	*Grammar*	*Punctuation*	*Spelling*
Purpose	100									
Audience	31	100								
Vocabulary	6	31	100							
Style	8	46	54	100						
Development	10	33	49	60	100					
Organization	15	40	44	59	60	100				
Sentences	3	28	39	50	39	44	100			
Grammar	2	18	28	26	19	20	48	100		
Punctuation	-1	15	25	21	15	19	54	46	100	
Spelling	8	21	25	23	17	21	39	52	44	100

($p < .001$)

Table 8. Rotated (Varimax) factor pattern (decimals omitted) for raw scores of students in Grades 4–6 on a test of written language (n = 848)

	Factors			
	I	*II*	*III*	*IV*
1. Purpose	-1	-1	1	92*
2. Audience	47*	19	2	59*
3. Vocabulary	72*	34	-1	-4
4. Style	82*	13	16	9
5. Development	82*	2	11	6
6. Organization	76*	-2	25	19
7. Sentences	43*	23	73*	2
8. Grammar	13	76*	34	0
9. Punctuation	5	34	83*	0
10. Spelling	10	84*	19	11
Variance explained by each factor	2.89	1.62	1.47	1.25
Final communality estimate				7.22

Note: Loadings greater than .40 are flagged with asterisks.

Each solution produces a new *factor structure,* or a matrix of loadings on factors. Each loading expresses the relationship between the variable on that row and the factor in that column. As Table 8 shows, the loading of each subtest (row) on each factor (column) reflects the strength of the relationship between the subtest and the factor. Although these loadings are actually less than 1.0, Table 8 uses the common convention of omitting the decimals (e.g., 19 instead of .19, as noted in the table caption). For convenience in evaluating the size of each loading in Table 8, all values greater than 40 (an arbitrary criterion) have been marked with an asterisk.

The factor in the first column to the right of the variable names always accounts for the greatest percentage of the variance of the variables. Each successive factor to the right accounts for a decreasing percentage of the variance. Each loading indicates the relationship of a variable with a factor. (The variables in the measure of writing reported in Table 8 are scores on specific elements of the written work called *features* by the test's authors.) Interpreting the loadings of each feature on each factor provides a look into what the test appears to measure.

In Table 8, it appears that four main patterns account for most of the variance in these features. The largest pattern is identified by the high loadings of vocabulary, style, development, and organization on Factor I. Two other features, audience and sentences, are also strongly related to this first factor. The second factor is identified by two features: grammar and spelling. The third factor is identified by sentences and punctuation, and the fourth factor is identified by purpose and audience (audience is also related to the first factor). Based on a logical consideration of these features, the test's authors identified the first factor as the writer's development of the work, the second factor as the writer's fluency with mechanics, the third factor as sentence structure, and the fourth factor as the writer's orientation to the reader.

This overall pattern of results may be described as a nearly "simple structure" because nearly all the subtests load high on one factor and low on the other factors. Although researchers generally use factor analysis to analyze intercorrelations of scores on subtests, larger samples permit analysis of more variables. For example, analyses of items within a subtest can be helpful in determining if one or more factors are required to explain the variance of the subtest.

Test users can learn much by reviewing factor analysis results in this way. That is, identify which subtests have high loadings and, for each subtest, note the highest loading on that factor. Then, look for any noted patterns in the loadings. Consider what content or task format might be shared by subtests that have high loadings on the same factor. Also examine subtests that logically seem to have much in common but that exhibit low loadings instead of high loadings. Try to identify differences in content, task formats, or other characteristics that might explain why these apparently similar subtests do not load together.

However, be wary of factor analysis tables that appear to be incomplete or that leave the loadings of some subtests unresolved. Be cautious about drawing conclusions unless each subtest shows a moderate to high loading on at least one factor. Unless the last factor in the table shows no high loadings, the author should report if the factor matrix could not be rotated further. If this information is not reported, try to predict what might be the loadings of the variables if still another factor could be generated.

Thus, in reviewing factor analysis results, always try to relate the patterns of loadings to the test maker's explanations of test content or to predictions stated in the test maker's theory or model. If logical or interpretable patterns can be found in the loadings (or in the correlations if no factor analysis is reported), they provide important evidence that the test fits an interpretable scheme. Factor analysis not only simplifies a large matrix of correlation coefficients, but also may help clarify how well the subtest scores relate to cluster scores or composite scores. By trying to relate these results to the test plan and to one's own knowledge of the domains tested, a user can decide how well these results align with the rationale for the ways test scores are to be interpreted.

As noted earlier, these various sources of statistical evidence are often referred to as *construct validity*, although construct validity more accurately includes both logical and empirical support for validity. Regardless of how sources of validity evidence are categorized, users should be wary of a test that fails to address most of the questions posed here. One can also look beyond the test manual to research articles, published critiques, or the experiences of others who have used it. The search for evidence of validity does not end in tables of numbers in the technical manual, because establishing validity is a matter of collecting evidence to support specific uses of a test. Messick (1989) has pointed out that:

> Over time, the existing validity evidence becomes enhanced (or contravened) by new findings, and projections of potential social consequences of testing become transformed by evidence of actual consequences and by changing social conditions. Inevitably, then, validity is an evolving property and validation is a continuing process. Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what the test scores mean. (p. 13)

QUESTIONS CONCERNING TEST RELIABILITY AND ITS IMPORTANCE

An axiom in measurement is that reliability is a necessary but not a sufficient condition for validity. This means that a test cannot be valid unless it is reliable, but reliability is not a guarantee of validity. Certainly, users need to know if a test will yield consistent results for a child, across different examiners or different administrations of the test. In fact, so much attention is given in technical sections to questions of reliability that Feldt and Brennan (1989) have noted that "the publication space accorded to reliability fails to reflect the widely accepted principle that the validity of a measure is a more crucial and comprehensive characteristic" (p. 143).

Unlike validity, which is a unitary notion (there are different kinds of evidence but not different kinds of validity), reliability is *not* unitary. That is, a test's reliability can be classified by the various sources of error possible: (a) items and subtests, (b) examiners, (c) conditions of time and place, and (d) test takers and standardization samples. Although these different types of reliability are usually considered separately, an approach called *generalizability theory* (Cronbach, Rajaratnam, & Gleser, 1963) provides a means to compute a coefficient that reflects more than one such source of error.

9. How different are the test items from each other?

Sometimes it is important to know how consistently the items in a test or subtest measure the same characteristic. Depending on the test, this may be a matter of *internal consistency*, or what Anastasi (1988) has called "inter-item consistency." The "split-half reliability" coefficient is an estimate of internal consistency based on splitting the test into two half-forms and obtaining the correlation between the two scores. (The resulting estimate must then be corrected to adjust the lower correlation between the two half-tests to reflect the length of the original test containing all the items.)

Different methods of splitting the items can produce different estimates of internal consistency. Splitting the items in many tests is done by assigning every other item to each half-form. For tests with items ordered by increasing difficulty, this has the advantage of balancing the difficulty of the two forms. Other aspects of the test's construction must be considered as well.

For example, consider two tests of listening comprehension. In one test, the student listens to a sentence or two and then answers a question. In the second test, the student listens to a longer narrative and then answers several questions concerning the story's elements. The first test might produce more consistency across items, because the answer to one item is not closely related to the answer to any other item. In the second test, however, the questions over each story could yield very different half-forms depending on how they related to such elements as the story grammar or influences of prior knowledge associated with the story.

These differences might produce different reliability coefficients depending on which items were assigned to which half-form. Unless the assignments balanced the items fairly well, a better method would be to compute the correlation called *coefficient alpha* (Cronbach, 1951). Because it is analogous to the average of all the possible corrected split-half correlations, coefficient alpha is an estimate of the lower bound of the test's internal consistency reliability.

The internal consistency coefficient is the estimate of reliability most commonly reported in test manuals, no doubt for several reasons: (a) it is easily computed, (b) it does not require administering the test twice to subjects and thus is not affected by intervening instruction or treatment, and (3) it does not require collecting additional samples beyond the standardization sample. The internal consistency coefficient is influenced by a test's length, and tests with more than 20–30 items generally have acceptably high coefficients even when the items appear to vary considerably in content and difficulty.

A useful caution in evaluating the reliability of an unfamiliar test is to avoid being satisfied with the internal

consistency coefficient (or any one coefficient, for that matter) and to look instead for estimates of reliability in terms of other contexts. For instance, one or more of the following questions should also be addressed.

10. How does performance on the test vary with different raters?

Test users generally expect that test results should depend only on the child's performance, not on the examiner's. Thus, a relevant question for many tests is whether a subject's performance is judged similarly by different examiners or raters. The *interrater reliability coefficient,* often used in reporting results of studies of different raters, is the correlation between scores obtained from two raters. However, be aware that this statistic may be misleading if not accompanied by a clear description of the study. For example, one may obtain a relatively high coefficient by correlating only the total scores of raters who actually exhibit low agreement on many individual items. Because it is also possible for raters to agree with each other by chance, one must look beyond the correlations for information on percentages of agreement among raters.

11. Does the interrater study consider accuracy as well as agreement?

Also check for details on the participants in the interrater studies. It is possible, for example, that two raters may exhibit high agreement by both producing inaccurate scores. Consider three raters, one who is an expert and two who are inexperienced or untrained. It is possible that the two novices might agree with each other but not with the expert. To conclude that the two novices exhibit high interrater reliability would not be a sound basis for certifying them as raters, nor would one want to train the expert to agree with the novices. In short, the novices' lack of agreement with the expert is not simply an issue of reliability. Accuracy, the more serious issue, is actually more an issue of validity than reliability.

12. How does performance on the test change over time?

The term *stability* describes the consistency of test and retest results of subjects whose relative performance has not changed. If a test yields the same results, it is considered a stable, or reliable, measure. However, if the test measures a trait one would expect to change with the passage of time, stability in the scores would be neither expected nor desirable. Thus, assessing stability involves comparing test and retest scores where development or learning has had little or no effect on the child's relative standing in the group.

When a retest should be expected to produce different score patterns, the issue is not as simple. For example, very stable scores for a group of subjects with varying severity of traumatic brain injuries might indicate that an instrument is too easy (or too difficult) for most of the highly changeable subjects to show much evidence of change. In this case, too much stability might be evidence of problems that call into question the validity of the test. Usually, the time between administrations is also relevant to the question. If the time interval in the example of the test of traumatic brain injury was several weeks, one might question the stability more than if the interval was a few days.

13. What confidence can one have that test scores are accurate?

Reports of test performance should acknowledge the imperfect reliability of the test due to different possible sources of error. An advantage of reliability estimates such as internal consistency, test–retest, and interrater coefficients is that they permit direct comparisons of different instruments or subtests. However, these coefficients are of limited value when interpreting scores for an individual child. A better approach is to express the measurement error as an adjustment to the reported score, that is, in terms of a margin of error or a "confidence range." Pollsters use this device when they report results as having a "± 3% error margin," for example. For the same reasons that this caveat has become a standard in reporting poll results, test users should also report an error margin for each test score.

The *standard error of measurement* (SEM) provides a useful means of reflecting a test's reliability by changing the focus from a single, absolute score point to a range around the obtained score (e.g., 108 plus or minus 3). In other words, the SEM reflects reliability in the same unit of measurement as the obtained score (typically raw scores or standard scores). This makes it easier to communicate results to parents, students, or other audiences unfamiliar with these measurement concepts.

One can also obtain a level of confidence for the range and compute different score ranges for different levels of confidence. The larger the error range, the more confident one can be that the student's "true score" lies within it. A score range based on 1 SEM above and below the obtained score allows one to be approximately 68% confident that the student's true score is enclosed by the range. A score range based on 2 SEMs allows one to be approximately 95% confident the true score is enclosed by the range.

14. How can I express measurement error in interpreting score differences on different subtests or clusters?

The SEM can help to avoid overinterpretation of differences between scores that might be caused by chance or measurement error. Hills (1981, pp. 247–252), among others, has suggested a way to implement this use of the SEM by plotting the scores and their confidence intervals on a simple graph, such as the one shown in Figure 1. The rule of thumb is this: Regard scores with overlapping confidence bands as too close to consider different. In Figure 1, the development and fluency scores on a measure

Figure 1. Standard scores plotted for two scores (13 and 11) with overlapping confidence ranges.

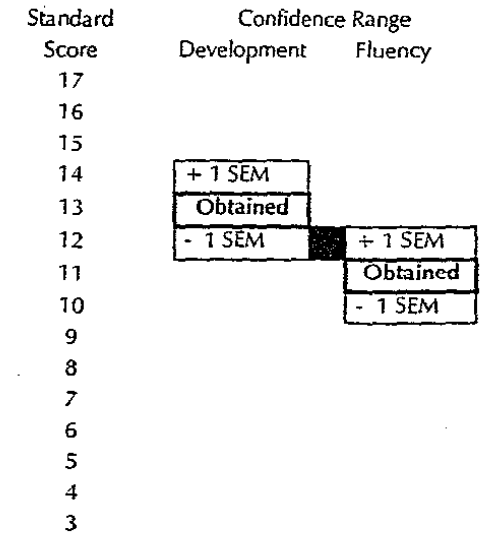
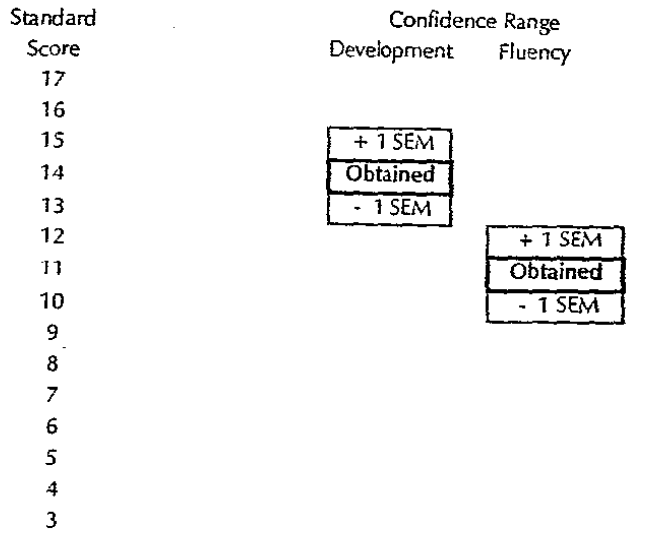


Figure 2. Standard scores plotted for two scores (14 and 11) with confidence ranges that do not overlap.



of written composition overlap. In Figure 2, the upper limit of the confidence range for fluency does not overlap the lower limit for development, so the examiner can be approximately 90% confident that the student's true score on development is higher than the true score on fluency.

A caution to remember is that different SEMs can be estimated based on different types of data (e.g., internal consistency, test–retest, and interrater). Therefore, check the manual to determine which reliability coefficients were used to compute the reported SEMs. Finally, note that the size of the SEM actually varies across the range from the lowest to highest score. However, the most common formula for the SEM produces a single value for all the scores on a test. Many test manuals use this approach, but some are technically more correct when they report SEMs of different magnitude for different raw scores.

QUESTIONS CONCERNING STANDARDIZATION DATA AND NORMS

Besides stating a clear rationale for the purpose and design of the test, the manual should also describe the steps taken to bring the test to its final form, including any field tests or tryouts that were conducted before standardization. The manual should also provide enough details concerning the standardization sample that a user can judge if the norms are appropriate for use with a particular child or types of children to be evaluated. Two important criteria are the reasonable representativeness and adequate size of the sample.

15. How well does the sample represent the population to which the child will be compared?

The answer to this question is important because the usefulness of norm-referenced interpretations depends on how relevant and reasonable a comparison can be made between the child and the sample. The representativeness of a norming sample is the extent to which the proportions of different groups in the sample match the proportions of those groups in the relevant population. However, the general population is not always the only, or even the most, important reference group for a subject.

For example, performance on a cognitive–linguistic test for individuals with traumatic brain injuries might be compared to the performance of subjects with similar injuries or subjects with no known injuries. In fact, both comparisons might be useful—the first to make decisions concerning interim support or intervention, and the second to make judgments regarding the child's readiness to return to home and school. Thus, representativeness may be a moving target because representative members of one reference group (subjects with traumatic brain injuries) may be very different from representative members of another reference group (subjects without traumatic brain injuries).

Keep in mind that representativeness of norms is not a guarantee for relevance or appropriateness, especially when the nature of a test calls into question significant differences within the population. Age norms based on a representative sample of the U.S. population may not be very relevant for assessing the expressive morphology of a bilingual child whose primary language is not English. Here, relevance is a more pressing concern than overall representativeness. A sample of the general population might include some percentage of bilingual speakers, or it might include only children whose primary language is English. In either case, the norms would likely reflect much lower performance than norms based on a sample of speakers of bilingual children for whom English was not the primary language. A score based on a sample of primary-English speakers might even be regarded as indicating a language deficit, although the more appropriate explanation would be the differences between the primary languages of the child and the children in the norms sample.

In some cases, a more appropriate comparison may be with local norms or special norms for specific groups obtained from ongoing administrations of the test. For example, one such group might consist of similar bilingual children whose second language is English regardless of their primary language, and another group might include only children with a particular first language and English as their second language. Computer software for constructing specialized norms sets such as these is available to test users (Sabers & Hutchinson, 1990).

16. Is the use of percentile ranks and standard scores clear, understandable, and appropriate?

The two most common methods of expressing normed scores are percentile ranks and standard scores. A *percentile rank* (PR) reflects the percentage of subjects in the sample who scored at or below a given raw score. A *standard score* is a score on a scale with equal intervals, such as those shown in Table 9. Although the different scales in Table 9 seem to reflect very different values, all are simply different expressions of how raw scores relate to the mean and standard deviation (SD) of the sample. However, the differences between percentile ranks and standard scores are important, and a test manual must make clear how these scores were developed and how they are to be interpreted.

PRs have the advantage of being simple to understand and explain to parents, students, teachers, or others unfamiliar with how to interpret test scores. A possible disadvantage of PRs is that inexperienced users can sometimes overinterpret PRs within the normal range (e.g., PR 16 and PR 84 seem far apart, but each is only 1SD from the mean).

Standard scores have the advantage of flexibility and precision in expressing performance. A score on one scale can even be directly translated into its equivalent on another scale, although with some loss in precision if converting from a more precise to less precise scale. As Table 9 shows, a composite score of 85 on the metric used by the Wechsler scales is equivalent to a score of 40 on a *T*-score scale. Although it would be unusual to report Wechsler scores on a *T*-score scale, it would not misrepresent the subject's performance, only express it less precisely than does 85 on the Wechsler metric. Standard scores also can be added to produce composite scores that reflect different combinations and weights of subtests. Most widely used language tests have a mean of 10 and a standard deviation of 3 for subtests and a mean of 100 and standard deviation of 15 for composite scores.

There are substantial differences between the appropriate uses of PRs and standard scores. Users (and a few test makers) sometimes forget that PRs, unlike standard scores, should not be added, subtracted, or averaged. Consider the following example, shown in Table 10. Scores are obtained on two subtests that both have a mean of 10 and SD of 3. The obtained standard scores, 17 and 7, have been averaged, yielding a result of 12 (17 + 7 = 24 divided by 2). In the test manual, the standard score of 17 on Subtest A is equivalent to PR 99, and the standard score of 7 on Subtest B is equivalent to PR 25. The average of PR 99 and PR 25 is 62 (99 + 25 = 124 divided by 2). However, in the test manual, the standard score of 12 is equivalent to PR 75, not the 62 obtained by averaging the PRs. Thus, because PRs do not represent a scale with equal intervals, it is inappropriate to average a child's PRs or to average the PRs of several children. Standard scores, rather than percentile ranks, should be used when averaging scores, whether the average is for scores of an individual or for scores obtained from several individuals.

17. Are the standard scores linear or normalized?

Users should also be aware that two ways of computing standard scores can result in different results, depending on how the distribution of raw scores differs from the familiar normal curve. In Table 11, columns 3 and 4 show substantial differences between these two types of standard scores for a set of Grade 10 norms on a writing mechanics subtest (Warden & Hutchinson, 1992). The scores in column 4 are *linear standard scores*, which were computed from the mean and standard deviation of the sample. (Specifically, each linear standard score is computed by multiplying the *z*-score for the corresponding raw score by 3 and adding 10.) The scores in column 3 are *normalized standard scores*, so called because they are based on the PRs obtained from the sample and the relationship between PRs and *z*-scores in a normal distribution. (Specifically, the normalized standard score uses the obtained PR for a given raw score and then reports the *z*-score equivalent to that PR

Table 9. Types of standard scores.

Score Name	*Mean*	*SD*	*Score at +2 SD*	*+1 SD*	*Mean*	*-1 SD*	*-2 SD*
z-score	0	1	2	1	0	-1	-2
T-score	50	10	70	60	50	40	30
Wechsler subtests	10	3	16	13	10	7	4
Wechsler composites	100	15	130	115	100	85	70
Stanford Binet subtests	50	8	66	58	50	42	34
Stanford Binet composites	100	16	132	116	100	84	68
NCE	50	21.06	92.1	71.1	50	29.9	8.8

Table 10. Standard score (SS) and equivalent percentile rank (PR) from a norms table.

	Average	
	SS	*PR*
Subtest A	17	99
Subtest B	7	25
Sum of A+B	24	124
Average (A+B/2)	12	62

Note: 75, not 62, is the PR of SS 12 in the norms table.

Table 11. Linear and normalized standard scores at Grade 10 for a test of Writing Mechanics with a mean of 12.1 and standard deviation of 2.6.

1. Raw score	*2. Percentile rank*	*3. Normalized M=10, SD=3*	*4. Linear M=10, SD=3*	*5. Linear z M=0, SD=1*
16	96	15	15	1.50
15	88	14	13	1.12
14	71	12	12	0.73
13	57	11	11	0.35
12	44	10	10	-0.04
11	31	9	9	-0.42
10	21	8	8	-0.81
9	17	7	6	-1.19
8	11	6	5	-1.58
7	0.6	5	4	-1.96
6	0.4	5	3	-2.35
5	0.2	4	2	-2.73
4	0.1	3	1	-3.12
3	0.1	3	-0	-3.50
2	0.1	3	-2	-3.88
1	0.1	3	-3	-4.27

in a normal distribution.) The more a distribution of raw scores differs from the normal distribution, the more the linear and normalized standard scores will differ (as do those in Table 11).

These differences are important because the normalized scores facilitate comparisons that the linear scores do not. For example, the Writing Mechanics subtest is accompanied by a companion subtest of Writing Development. Although it was normed on the same group of tenth graders, its mean and standard deviation are different from the mean and standard deviation of the Mechanics subtest. In other words, the scale for the Mechanics subtest is not the same as the scale for the Development subtest. Each scale is made up of equal intervals, but it is as if one scale is expressed in meters and the other in yards, and the two cannot be compared without some additional conversion. However, using normalized standard scores permits us to add, average, or directly compare the two scores because both scales use the same unit of measurement.

18. How well does the test sample behavior at the extremes?

Many tests attempt to sample content that is typical of subjects at relevant ages, grades, or other levels of development or ability. However, tests designed in this way often present very few opportunities for the subjects at either extreme—the very low scoring or the very high scoring subjects. Thus, a relevant question in reviewing a test is whether it has high *floors* or low *ceilings*. In the first case, the test may have so few low-level (easy) items that it is difficult to differentiate among subjects whose scores are very low. In the second case, the test may have so few high-level (difficult) items that it cannot differentiate among students whose scores are very high. On many clinical batteries designed to assess a wide range of ages or ability levels, this problem tends to appear only at the youngest and oldest ages or grades, because there are too few items easy enough for the youngest students or hard enough for the oldest students. However, this problem of test content is often confused with another problem, which is the inaccuracy of the norms at the extremes.

19. How well do the norms represent performance at the extremes?

Norms collected on representative samples generally have more subjects earning scores closer to the middle of the score range than at the extremes (very low or very high scores). In other words, more subjects earn the average score (and the scores closer to the average) than earn the extreme scores. Stated another way, the scores obtained from normal distributions do not provide as many cases to estimate the scores at the extremes as at the middle. Ironically, many clinical tests are intended for subjects whose performance is at one of the extremes—very able or very disabled subjects. Therefore, some test makers attempt to improve accuracy for the extremes in one of two ways. One is to "oversample the tails" by collecting more subjects at one or both extremes, for the express purpose of better representing these subjects. For example, to assess the mean length of utterance of children from ages 2 to 5, a test maker may elect to sample children from 1½ to 2 years of age. This larger sample of children below age 2 would provide a much sounder basis for estimating the range of scores of those at age 2. Another way is to "extrapolate" or estimate very extreme scores based on extending the downward trend of decreasing scores (at the lower extreme) or the upward trend of increasing scores (at the higher extreme). Either extrapolation or oversampling can provide greater accuracy in estimating norms for extreme scores, but the two do not necessarily yield the same results. Thus, the test maker should indicate which of these approaches, if any, was used to estimate norms at the extremes.

20. How does the test maker treat developmental norms such as age or grade equivalents?

Norms on tests that measure developmental improvement generally report standard scores or PRs for several different ages or grades. However, some of these tests also report scores based on the entire continuum from earliest to latest levels of development, such as age equivalents (AEs) from

5:0 (years:months) to 18:0 or grade equivalents from K:1 to 12:9. Although these developmental scores often provide a quick estimate of the age or grade where the subject's score was the average score, they do not have the same characteristics as a PR or standard score for the subject's specific age or grade.

Table 12 illustrates that an age equivalent score is simply the level at which a given raw score was the average for an age group. In some cases, such as AE 5.2, 5.3, and 5.5 in the example in Table 12, an age equivalent may be a calculated distance between age equivalents obtained from averages for a few groups. Thus, an age or grade equivalent does not reflect a distribution of scores at each month of a year-long age interval or grade level, except to indicate whatever raw score was the median (at the 50th percentile) at that interval. Because grade equivalents are typically constructed from samples collected at one or two times during the school year, the intermediate grade points must be calculated from these empirical points.

Table 12. Portion of an age equivalent scale for subjects tested by 3-month intervals.

Raw score	*Age equivalent*	*Explanation*
15	5:1	15 was the average raw score obtained for the group of subjects whose average age was 5:1 (those between ages 5:0 and 5:2).
16	5:2	16 is one-third of the difference between raw score 15 and raw score 18; 5:2 is one-third of the distance between AE 5:1 and AE 5:4.
17	5:3	17 is two-thirds of the difference between raw score 15 and raw score 18; 5:3 is two-thirds of the distance between AE 5:1 and AE 5:4.
18	5:4	18 was the average raw score obtained for the group of subjects whose average age was 5:4 (those between ages 5:3 and 5:5).
19	5:5	19 is one-half of the difference between raw score 18 and raw score 20; 5:5 is less than one-half and 5:6 is more than one-half of the distance between AE 5:4 and AE 5:7; the AE of 5:5 was assigned to raw score 19.
	5:6	There is no raw score for AE 5:6 because raw score 19 is AE 5:5 and raw score 20 is AE 5:7.
20	5:7	20 was the average raw score obtained for the group of subjects whose average age was 5:7 (those between ages 5:6 and 5:8).

Note that these developmental scores cannot provide the same information concerning a child's standing as that provided by standard scores or percentile ranks, which represent standing within a group of peers at the child's age or grade. More important, the measurement error associated with a raw score is often overlooked when age or grade equivalents are reported. For example, consider a 100-item test for which developmental scores have been computed across 10 years, from age 6 to age 15. And, assume that the 10 1-year distributions of test scores in the norming sample progressed so regularly that each of the 100 raw scores represented a different age equivalent. Finally, assume that the test's reliability was uniformly high across the 10 years, and the standard error of measurement (SEM) was only 2 raw score points at every score level.

Even with this powerful combination of desirable properties, Figures 3 and 4 indicate that one should interpret differences between age (or grade) equivalents with extreme caution. Using the SEM to build confidence bands around two obtained raw scores helps to illustate the error associated with age or grade equivalents. We begin by using a confidence level of 68% (1 SEM). On our fictional test, 1 SEM is a range of four raw score points (two above and two below the obtained raw score). Thus, the 4-point difference between a raw score of 34 (AE 8:4) for the first administration and 38 (AE 8:8) for the second administration should not be interpreted as significantly different, as indicated by the overlapping confidence bands in Figure 3. At the 95% level, a difference of 7 months (AE 8:4 to AE 9:1) could be attributed to measurement error, as illustrated by the overlapping confidence bands in Figure 4. For these and other reasons, test experts in many fields, including speech and language (McCauley & Swisher, 1984), have advised against the use of age equivalents.

CONCLUSION

Too often, users of tests regard the technical material reviewed here as information for someone else, put in manuals to fulfill requirements imposed by professional standards committees or to keep psychometricians employed. In fact, this information can provide a more thorough understanding of: (a) the test maker's reasons and assumptions in creating the test, (b) the information collected by the test maker to support the content and structure of the test, and (c) the methods and outcomes of studies of the test in use. This information can also give a user greater assurance when making inferences based on sound data and can stimulate greater cautions about making inferences without regard to measurement error. Taking these concerns into account, and acknowledging that all test results are bound by the contexts in which they were obtained, can actually inform and improve judgments about the clinical uses of a test.

Figure 3. Age equivalents and 68% confidence intervals plotted for raw scores with a standard error of measurement (SEM) of 2.

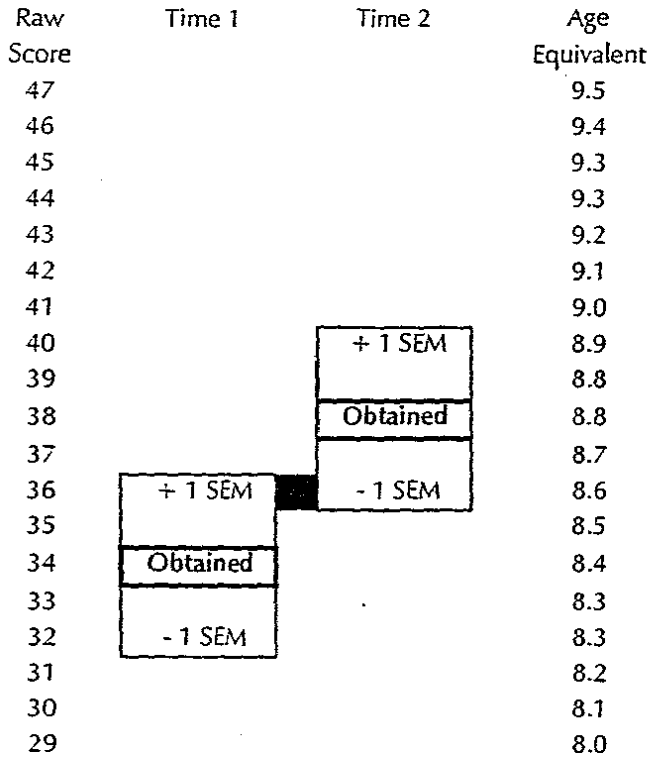
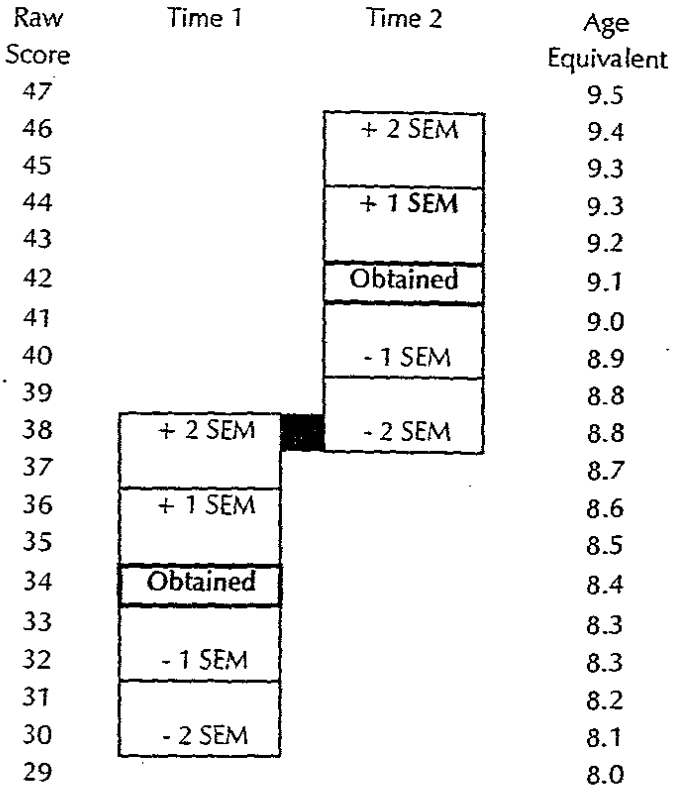


Figure 4. Age equivalents and 95% confidence intervals plotted for raw scores with a standard error of measurement (SEM) of 2.



ACKNOWLEDGEMENT

I thank Darrell Sabers and Rhia Roberts for comments on a previous draft of this manuscript.

REFERENCES

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137–163.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R.L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education/Macmillan.

Hills, J. R. (1981). *Measurement and evaluation in the classroom* (2nd ed.). Columbus, OH: Merrill.

Kretschmer, R. R., & Kretschmer, L. W. (1978). *Language development and intervention with the hearing impaired.* Baltimore, MD: University Park Press.

McCauley, R. J., & Swisher, L. (1984). Use and misuse of norm referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49,* 338–348.

Messick, S. L. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Sabers, D., & Hutchinson, T. A. (1990). *User norms software.* Chicago: Riverside Publishing.

Warden, M. R., & Hutchinson, T. A. (1992). *Writing Process Test.* Chicago: Riverside Publishing.

Received June 1, 1994
Accepted April 13, 1995

Contact author: Thomas Hutchinson, Applied Symbolix, 16 W. Erie Street, Suite 300, Chicago, IL 60610.