Familiar Strangers: Criterion-Referenced Measures in Communication Disorders

Rebecca J. McCauley
University of Vermont, Burlington

It is the third week of school and you have scheduled your first session with George, a 7-year-old boy who is meeting with little success in the early stages of learning to read following a history of language impairment including severe phonologic difficulties dating from preschool. Studying his file, there is no question in your mind that almost any of the norm-referenced tests you might use to reevaluate his speech and language skills will confirm his overall difference from his peers. In fact, you believe many of these tests will provide you with insight into some of the strengths and weaknesses of his communication.

At least as urgent a need as the continuing documentation of his problems, however, is your desire to know what gains he has made on the behavioral objectives targeted when he was last seen and where to begin your efforts to help him. The questions that press their way to the front of your mind are myriad: What is his intelligibility in the classroom? Which grammatical morphemes continue to be missing from his expressive language? What simplification processes characterize his speech errors? What is his understanding of vocabulary that will be used in his second-grade class? What understanding does he have concerning the sound structure of words and syllables that he can draw on for reading?

Fortunately, as you contemplate the clinical questions posed by George, you will have methods in mind to answer them, many of them involving the use of probes of your own construction. Despite their pervasiveness in your repertoire of clinical methods, and despite the skill with which you use them, you may only vaguely be aware that such measures can be described as “criterion-referenced,” and even less aware of a body of research and writing designed to aid you in their development and use.

Table 1 provides specific examples of a variety of such measures for school-age children. Although many of these measures are not referred to by their proponents as “criterion-referenced,” and although several of the measures listed would almost certainly fail to fit within more narrow definitions of the term or be considered well-developed criterion-referenced measures (cf. Nitko, 1983), all of these measures are “criterion-referenced” in that they are frequently interpreted in terms of established performance levels. One purpose of this article, therefore, is to call to the reader’s attention some of the basic properties of such measures—“familiar strangers”—that one may use daily yet have little opportunity to learn about.

Criterion-referenced measures are described as “familiar” here because they are probably more commonly used by speech-language pathologists than almost any other type of measure (Muma, Pierce, & Muma, 1983). Despite their ubiquity, however, such measures may also be the least understood of those we use—hence the label “strangers” to convey the sense of their mystery. Because norm-referenced tests have historically garnered more attention (e.g., Brown, 1989; McCauley & Swisher, 1984a, 1984b), the importance of criterion-referenced measures and principles guiding their use frequently have been unappreciated. In this article, therefore, the intent is to review the basic characteristics of criterion-referenced measures (especially as these contrast with norm-referenced measures) and outline a set of

ABSTRACT: Although frequently used in the assessment and treatment of communication disorders, criterion-referenced measures are often not well understood, making them both familiar and alien—thus, familiar strangers. This article is designed to better acquaint test users with the characteristics associated with the use and evaluation of criterion-referenced measures, particularly as they differ from norm-referenced measures. Guidelines are proposed for the evaluation and selection of standardized criterion-referenced measures as well as for the development and ongoing evaluation of informal criterion-referenced measures.

KEY WORDS: criterion-referenced measurement, informal measures, test development, test evaluation

122 LANGUAGE, SPEECH, AND HEARING SERVICES IN SCHOOLS • Vol. 27 • 0161-1461/96/2702-0122 © American Speech-Language-Hearing Association
Table 1. Examples from a variety of developmental communication disorders of measures that fit within a broad definition of criterion-referencing.

<table>
<thead>
<tr>
<th>Communication disorder</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hearing impairment</td>
<td>• Pure tone threshold testing (Carhart &amp; Jerger, 1959)</td>
</tr>
<tr>
<td></td>
<td>• Speech Recognition Threshold (American Speech-Language-Hearing</td>
</tr>
<tr>
<td></td>
<td>Association, 1988)</td>
</tr>
<tr>
<td></td>
<td>• Phonetic Level Speech Evaluation (Ling, 1976)</td>
</tr>
<tr>
<td>Fluency disorders</td>
<td>• Percentage of syllables stuttered</td>
</tr>
<tr>
<td></td>
<td>• Proportion of disfluency types</td>
</tr>
<tr>
<td></td>
<td>• Stuttering Severity Index for</td>
</tr>
<tr>
<td></td>
<td>Children and Adults (SSI-3) (Riley, 1994)</td>
</tr>
<tr>
<td>Voice disorders</td>
<td>• Maximum phonation time (Colton &amp; Casper, 1990)</td>
</tr>
<tr>
<td></td>
<td>• Vocal intensity (Colton &amp; Casper, 1990)</td>
</tr>
<tr>
<td></td>
<td>• Mean fundamental frequency (Colton &amp; Casper, 1990)</td>
</tr>
<tr>
<td>Phonological disorders</td>
<td>• Profile of phonological process use (Khan &amp; Lewis, 1986)</td>
</tr>
<tr>
<td></td>
<td>• Percentage of consonants correct</td>
</tr>
<tr>
<td></td>
<td>• Percentage of intelligible words</td>
</tr>
<tr>
<td></td>
<td>(Weiss, Gordon, &amp; Lillywhite, 1987)</td>
</tr>
<tr>
<td>Language disorders and language-learning</td>
<td>• Language sample analyses (Lahey, 1988)</td>
</tr>
<tr>
<td>disabilities</td>
<td>• Mean length of utterance (Brown, 1973)</td>
</tr>
<tr>
<td></td>
<td>• Decoding Skills Test (Richardson &amp; DiBenedetto, 1985)</td>
</tr>
</tbody>
</table>

guidelines for clinicians to keep in mind as they use such measures.

BACKGROUND

Historically, some of the obscurity surrounding criterion-referenced tests undoubtedly stems from their recent evolution—the term criterion-referencing having been coined only a little more than 30 years ago (Glaser, 1963; Glaser & Klaus, 1962). At that time, the idea of an alternative to norm-referencing grew out of the frustration test users experienced with the adequacy of norm-referenced tests for some purposes. Although such measures were often satisfactory for the purpose of identifying children as sufficiently different from their peers to require help in a particular area, they were far less useful for the purpose of planning the nature of that help. In particular, when it was realized that achievement tests might ideally reflect the extent to which an individual had mastered a particular body of knowledge or achieved a particular level of skill attainment (Glaser, 1963), theorists began to consider modifications in test development and interpretation that were, at times, at odds with procedures recommended for norm-referenced instruments (Carver, 1974; Gronlund, 1982). In this new approach to test development, for example, the performances of groups of individuals who were like those for whom the test was developed remained of interest as a means of establishing performance standards, rather than as a group of scores against which future test takers were to be compared (as would be the case if norm-referencing were used).

In the 1970s and '80s, there was a proliferation of criterion-referenced tests and of scholarly activity devoted to the implementation of criterion-referencing, particularly in educational applications (Nikko, 1983). Much of this work, however, took place independently under a variety of names, including domain-referenced testing, objectives-referenced testing, competency-based testing, and mastery testing (Berk, 1984a). In 1985, the influential Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council of Measurement in Education [AERA, APA, & NCME], 1985) recognized the growing stature of such tests, which were defined as follows: “A criterion-referenced test. A test that allows its users to make score interpretations in relation to a functional performance standard, as distinguished from those interpretations that are made in relation to the performance of others” (p. 90).

The Standards also recognized “domain-referenced tests,” where a domain-referenced test was defined as “a test that allows users to estimate the amount of a specified content domain that an individual has learned. Domains may be based on sets of instructional objectives, for example” (AERA, APA, & NCME, p. 91). For many authors, however, the distinction between these closely related concepts is not maintained, and criterion-referenced testing is used as the cover term. In the interests of simplicity, it will be used that way here.

In the late 1980s and early '90s, it appears that “criterion-referencing” has been superseded as the compelling cutting-edge of educational and psychological testing. In particular, developments such as computerized interactive and adaptive testing, as well as portfolio assessment, appear to have assumed that position. Nevertheless, criterion-referencing continues to receive attention as an invaluable mode of test score interpretation (e.g., Cizek, 1993; Millman & Greene, 1989; Salvia & Ysseldyke, 1991).

Currently, in the field of speech-language pathology, the value of criterion-referencing is frequently affirmed in texts dealing with diagnostic methods (Haynes, Pindzola, & Emerick, 1992; Newhoff & Leonard, 1983) and in texts describing assessment and intervention for specific categories of communication disorder (Lahey, 1988; Nelson, 1993; Reed, 1994; Wallach & Butler, 1994). These sources generally recommend that criterion-referenced measures be used in two instances: (a) when the norms that might be used to provide a norm-referenced interpretation of test performance are unavailable or inappropriate, or (b) when information concerning specific skills or behaviors of the test taker is required by the clinical question being posed.

The first of these two instances occurs frequently, when one tests a student who speaks a dialect or has a cultural
background that differs from that of the normative sample (Taylor & Anderson, 1988; Terrell, Arensberg, & Rosa, 1992). It also occurs when one tests a student with mental retardation (Snyder-McLean & McLean, 1988), because neither age nor “mental age” can be expected to yield a meaningful basis for selecting a truly relevant normative group for such children (Salvia & Ysseldyke, 1991).

The second instance in which criterion-referenced measurement tends to be recommended occurs when specific information concerning behaviors or knowledge is required—as it is at many points during the natural course of a clinical relationship. During diagnosis of a developmental communication disorder, for example, an appropriate criterion-referenced measure can help support a specific diagnosis or provide a focus for treatment planning robust enough to guide the selection of treatment tasks, stimuli, and settings (Hutchinson, 1988; McAuley & Swisher, 1984b). Similarly, as treatment progresses, information concerning specific patterns of skill or knowledge acquisition and generalization will facilitate planning by highlighting areas in which change is or is not occurring through the use of criterion-referenced pre- and post-testing. Recently, Tomblin (1994) proposed the term “client-referenced” to refer to clinical measures in which the criterion relates to the client’s own past performance on the measure of interest. For the purpose of examining changes over time, criterion-referencing is applied frequently in the form of the assessment component of behavioral objectives.

ESSENTIAL CHARACTERISTICS OF CRITERION-REFERENCED MEASURES

As is evident from Table 1, both formal, or standardized, tests and informal clinician-developed measures can be criterion-referenced. In fact, currently in speech-language pathology, criterion-referencing is still far more common in informal than in formal measures. Returning to the case of George, for example, at least two of the five questions raised concerning his communication (i.e., those dealing with intelligibility and classroom vocabulary) would require the use of an informal criterion-referenced measure. Use of informal criterion-referenced measures places an especially great burden on the clinician to consider their adequacy because they will almost always lack the documentation typically supplied by test developers for standardized measures.

Another point that should be evident from examination of this table is that some of the measures listed as criterion-referenced can also be used with norms and thus be considered norm-referenced as well. For example, if the Khan-Lewis Phonological Analysis (Khan & Lewis, 1986) were to be used to address George’s use of simplification processes, his overall score could first be compared against the performance of a group of same-age peers to confirm that he has a phonological impairment (a norm-referenced interpretation of the test performance) and then could be used to plan treatment when the most-frequently occurring processes are selected as targets (a criterion-referenced interpretation of test performance). In fact, norm- and criterion-referencing are often and probably best considered as modes of interpretation rather than mutually-exclusive categories of tests (AERA, APA, NCME, 1985; Nitko, 1989). However, it is rare that a single instrument can support both modes of interpretation with equal validity. Should a test developer attempt to devise an instrument with such capabilities, the volume of evidence required to suggest validity for both purposes would be immense.

A comparison of norm- and criterion-referencing for standardized measures may prove helpful in highlighting the specific features that make an instrument that is good as a criterion-referenced measure less likely to be good as a norm-referenced measure and vice versa. Table 2 summarizes four features of criterion-versus norm-referenced measures that are often described in discussions of this distinction (Carver, 1974; Gronlund, 1982).

First, whereas norm-referenced measures are designed to rank individuals in terms of relative performance, criterion-referenced measures are designed to distinguish between performance levels through the specification of knowledge the test taker has or of tasks the test taker is able to perform successfully. This difference leads to somewhat incompatible bases for the selection of items and scores used to summarize test performance. It also leads to the comparison of a test taker’s test score, not to a group of scores obtained by a normative sample, as is the case in norm-referenced interpretations, but to a specific score or set of scores, which are alternatively called cutoffs, criterion levels, or performance standards.

Selection of a specific score or scores to differentiate performance levels, for example between those scores seen as adequate or indicative of “mastery” from those seen as inadequate or indicative of “nonmastery,” is one of the thorniest problems facing developers of criterion-referenced instruments (Berk, 1984a; Cizek, 1993; Nitko, 1983; Shepard, 1984). It also presents a point of possible confusion because test developers often use empirical information concerning group performance to guide their selection of cutoffs, thereby blurring the concepts of norms versus standards unless it is understood that norm-referencing occurs when an individual’s performance is primarily interpreted in relation to the performance of others.

Table 2. Comparison of norm- and criterion-referencing.

<table>
<thead>
<tr>
<th>Norm-referencing</th>
<th>Criterion-referencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental purpose is to rank individuals</td>
<td>Fundamental purpose is to distinguish specific levels of performance</td>
</tr>
<tr>
<td>Test planning addresses a broad content</td>
<td>Test planning addresses a clearly specified domain</td>
</tr>
<tr>
<td>Items are chosen to distinguish among individuals</td>
<td>Items are chosen to cover content domain</td>
</tr>
<tr>
<td>Performance can be summarized meaningfully using percentile or standard scores</td>
<td>Performance can be summarized meaningfully using raw scores</td>
</tr>
</tbody>
</table>
Interpretation of pure-tone threshold testing provides a good example of a criterion-referenced measure for which standards were established using normative data. Although normative populations were used to establish the standards on which pure-tone testing is based, the way an individual's performance is interpreted is not through direct reference to the hearing performance of others (e.g., John scores at the 30th percentile for hearing at 2,000 Hz), but through reference to a standard (e.g., John shows a hearing sensitivity of 50 HL at 2,000 Hz).

Second, whereas norm-referenced tests usually contain items intended to sample broadly from the skill, ability, or behavior being assessed, criterion-referenced tests contain items that more thoroughly cover a smaller domain. This means that from the first stages of item preparation, test content is chosen in a way that samples more or less narrowly from the domain being tested.

There are numerous schemes for moving from the inferences to be made to decisions concerning the way in which items will be posed, the nature of expected responses, and the way in which they will be evaluated (Nikitko, 1983; Popham, 1984). Many of them involve the development of very detailed specifications before the creation of items to ensure thorough coverage of the skill area to be tested, with the level of detail required falling well above that required for norm-referenced measures.

For example, in the item plan guiding the development of just one subtest, the authors of the Decoding Skills Test (Richardson & DI Benedetto, 1985) developed items that varied all of the following characteristics: mono- versus polysyllabic structure, single versus multiple consonants, short versus long vowels versus vowel digraphs, and real versus nonsense words. From this example, it should be evident that attempts to produce a criterion-referenced measure for language performance will necessarily be focused on only a very small area of language, or will need to be incredibly, perhaps impossibly, lengthy.

Third, test items in norm-referenced tests are typically chosen on the basis of their ability to discriminate among individuals, whereas test items in criterion-referenced tests are chosen to discriminate differences in performance. For both types of tests during item tryout, items are administered to a group of test takers who are similar to those for whom the test is intended. Then, during item analysis, the group's performance on each item is scrutinized.

For a norm-referenced test, the test developer would choose for the final version of the test those items on which test takers performed variably. Thus, any items that were passed by all test takers or failed by all test takers would be rejected because they did not discriminate among individuals. In contrast, for a criterion-referenced test, the test developer would choose for the final version of the test those items that were always passed by individuals (called masters) who had acquired, or mastered, a particular skill and always failed by those (called non masters) who had not mastered that same skill. This choice would be made regardless of how many test takers in the item tryout had failed or passed the individual item.

An example of item selection for a criterion-referenced instrument can be drawn from the Oral-Speech Mechanism Screening Examination (St. Louis & Ruscello, 1987), in which numerous items were included despite preliminary data suggesting that almost all of the pilot group passed those items. The inclusion of those items was justified by their importance to a complete picture of the unimpaired function of the oral-speech mechanism.

Finally, because the score we use to summarize a performance reflects the way in which we are interpreting it, norm-referenced tests more often use types of scores that reflect the test taker's score relative to those of other test takers—namely, percentiles, standard scores, and so forth. In contrast, criterion-referenced tests frequently make use of raw scores or percentages that are compared to a specific cutoff or criterion level—both for formal and informal measures.

In communication disorders, decisions regarding behavioral objectives, which are often stated in terms of informal criteria of speech or language behavior, serve as excellent examples of the use of raw scores or percentages as cutoff scores for decision making. Often, such cutoff scores are determined by the arbitrary selection of a level that is viewed as reasonably, but not overly, strict—80%, for example (Eger, Mient, & Cushman, 1986). This approach has been termed the "feet on the desk" procedure (Nikitko, 1983) because the selection process is guided by the kind of cogitation that can occur with one's feet comfortably planted on the desk, rather than that occurring after the brow-moistening accumulation of empirical evidence as input to the decision-making process.

In summary, then, although criterion-referenced and norm-referenced measures share some important commonalities, they present distinct challenges to their developers and subsequent users. Whereas much has been written explaining the means by which norm-referenced test developers and test purchasers can better ensure the quality of such measures, little comparable guidance has been provided for criterion-referenced measures in the field of communication disorders. In the next two sections of this article, guidelines are offered for the review and development of criterion-referenced measures.

GUIDELINES FOR STANDARDIZED CRITERION-REFERENCED MEASURES

Despite the significant differences between norm-referenced and criterion-referenced measures described above, the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985) provides guidelines, or standards, for test development and use that encompass both types of measures. Most standards are stated in a general manner, thereby permitting differences in technical implementation, depending on the type of measure. Others are clearly more relevant for one of the two types of measures. The intermingling of standards for both types is a tangible outgrowth of the Standards' focus on particular instances of test use and interpretation as the point at which decisions regarding psychometric adequacy are ultimately relevant (Messick, 1989).
The specific guidelines offered in Table 3 were developed with the intent of preserving the Standards' unitary approach to measurement while acknowledging the unique and emerging methods used in the development of standardized criterion-referenced instruments. They represent only a limited subset of characteristics thought to be important to the evaluation of criterion-referenced measures. (Readers are encouraged to refer to the Standards [AERA, APA, & NCME, 1985], Berk [1984b], and Nitko [1983] for more comprehensive guidelines.)

In this section, each guideline is offered in terms of relevant recommendations from the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985), limitations resulting when deficiencies are noted, and means by which the test user can recognize a test's fulfillment of the guideline. In addition, ways in which these guidelines differ from those relevant to tests supporting norm-referenced interpretations will be described.

(1) Clear Definition of the Test Domain

The specifications used in constructing items or selecting observations and in designing the test instrument as a whole should be stated clearly. The definition of a universe or domain that is used for constructing or selecting items should be described. (AERA, APA, & NCME, 1985, p. 25)

Because the meaningfulness of a criterion-referenced interpretation rests on the comparison of a sample of the test taker's behaviors to a specified standard of performance, it is crucial that the test user be aware not only of how broadly, but also how broadly or narrowly, the content domain has been defined during the test's development. Without detailed knowledge of how items were selected or written, the test user may misconstrue the test taker's competencies and thus undermine the value of the test's use. Thus, for example, the author of a measure designed to examine phonological process use in a child's speech should include an account of the possible frequency of occurrence of a given process for the stimuli used, as well as an account of the phonetic context in which each process was probed. This information should be provided because both factors are known to influence apparent patterns of process use (McReynolds & Elbert, 1981; Stoel-Gammon & Dunn, 1985).

In order to be assured that the test domain has been clearly defined, the test user can look for a clear description of the area of knowledge or behavior being assessed. That description should state the tasks to be completed, the plan used to guide item construction, and the methods used to evaluate the appropriateness of included items. As part of a suitable plan, users can expect to find clear delineations of whether items are interchangeable or are organized along a continuum based on difficulty or complexity, degree of proficiency, or level of development (Nitko, 1983).

With the increasing integration of language with curricular concerns for school-age clients, curriculum-based language assessment has been suggested as providing the framework for the specification of informal, often observational, measures (e.g., Nelson, 1989). Regardless of their source, clear item specifications enable a test's user to decide about the appropriateness of the measure for use in answering a specific clinical question concerning a given client.

Information concerning test content and methods in item development and selection are also relevant when norm-referenced interpretations of test performance are planned. However, the need for more detailed delineation of how the content domain was conceptualized and sampled is heightened for a test that will be used to choose treatment materials or methods of intervention.

(2) Evidence of Validity for the Test User's Intended Application

"Evidence of validity should be presented for the major types of inferences for which the use of a test is recommended. A rationale should be provided to support the particular mix of evidence presented for the intended uses" (AERA, APA, & NCME, 1985, p. 13).

When a test taker's performance is to be interpreted in terms of having met or not met a particular performance standard, the ultimate concern is that of validity, where validity refers to the ability of a measure to be used meaningfully for a given purpose. Evidence of validity has traditionally been marshaled under the headings of content, criterion-related, and construct validity, which correspond to overlapping (rather than alternative) empirical and logical approaches taken during a measure's development and use.

Because the ultimate determination of validity must be made with the testing purpose and test taker in mind, what constitutes adequate evidence will vary considerably with variations in testing purposes and test takers. For example, evidence supporting the use of a measure to help identify where to begin treatment for a child with difficulties in phonological processing will be quite different from that supporting the use of the same measure to help identify that a child is likely to possess precursory skills for learning to read. Fuller discussions of the complexity involved in validity assessments for criterion-referenced measures are available in Hambleton (1984).

Some of the types of evidence suggesting validity for the use of a criterion-referenced measure can be illustrated using the hypothetical example of a criterion-referenced language scale contemplated for use in the description of language skills of a young student who speaks a nonstandard dialect. In such a case, the description of the test's design and development under the first guideline offered above would be particularly important. In addition.

---

**Table 3. Guidelines for the evaluation of standardized criterion-referenced measures.**

<table>
<thead>
<tr>
<th>Guideline number</th>
<th>Guideline content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Clear definition of the test domain</td>
</tr>
<tr>
<td>2</td>
<td>Evidence of validity</td>
</tr>
<tr>
<td>3</td>
<td>Evidence of reliability</td>
</tr>
<tr>
<td>4</td>
<td>Careful description of the test takers used in studies of reliability and validity</td>
</tr>
<tr>
<td>5</td>
<td>Detailed description of test administration</td>
</tr>
<tr>
<td>6</td>
<td>Detailed description of test user qualifications</td>
</tr>
</tbody>
</table>
evidence that experts in the domain of interest guided or agreed with the inclusion and ordering of specific tasks would provide further evidence that the measure would be likely to function as intended, that is, that it has content validity for this particular application. Although it is often claimed that such information is sufficient to document validity for a criterion-referenced instrument, other types of information are indeed quite pertinent to that conclusion (Hambleton, 1984).

Additional sources of relevant validity evidence for the hypothetical case described above would be the results of studies showing that children for whom there were no concerns regarding language were successful in performing tasks at their age level, or that children who failed at their age level were unable to respond correctly to the tested content using other, validated measures. Such studies may be conducted before item analysis during the test’s development (Berk, 1984a), during studies designed to establish performance standards (Shepard, 1984), or after the test’s completion during validation studies (Hambleton, 1984). Also crucial for this application of the language measure would be data showing that passing performances were obtained by children who are known to have normal language skills and who are from the same minority group as the child to be tested.

Readers will recognize the kinds of evidence described above as quite similar to those discussed for norm-referenced interpretations (e.g., McCauley & Swisher, 1984a). One difference, however, that may also be evident is that, whereas the decision to be defended in norm-referenced interpretation is one of relative ranking compared to a particular normative group, the decision to be defended in criterion-referenced interpretations is often a dichotomous decision, often a “pass-fail” or “mastery-nonmastery” decision as to whether an individual’s performance met or did not meet a specified performance level. Thus, in order for validity to be supported, the specific criterion level used as a cutoff as well as the measure as a whole must be positively evaluated (Shepard, 1984).

(3) Evidence of Reliability

“Typically test developers and publishers have primary responsibility for obtaining and reporting evidence concerning reliability and errors of measurement adequate for intended purposes” (AERA, APA, & NCME, 1985, p. 19).

In the evaluation of reliability for a criterion-referenced measure or interpretation, the desired characteristic can take a variety of forms: consistency in decision outcome (e.g., master/nonmastery), consistency in distance from the cutoff score, or consistency in obtained scores when closely related versions of a measure are administered (Berk, 1984a). Which of these forms is seen as most appropriate will depend on the application to which the measure will be put and the orientation of the test author. Measures used to describe observed patterns of consistency may be labeled agreement indices or reliability coefficients and are almost overwhelming in number (Berk, 1984a; Feldt & Brennan, 1989; Subkoviak, 1984).

Without information regarding consistency, the test user may mistakenly classify an individual as requiring further attention in a specific area or, conversely, as having mastered a specific area when those inferences are not valid. For speech-language pathologists, this may result in the selection of inappropriate treatment objectives or inappropriate timing of movement along a continuum of tasks or stimuli.

As an example, consider the use of a criterion-referenced measure to determine whether or not a child possesses the phonological processing skills required in early reading acquisition. The test user would want to know that the child’s passing or failing such a measure would be consistent regardless of the exact items used, the identity of the evaluator, and so on. Evidence that relevant groups of children were consistently classified by the measure would help provide that assurance. However, given the still emerging nature of criterion-referenced measures relevant to communication disorders, such evidence may not be offered. Instead, reliability coefficients based on methods traditionally used for norm-referenced measures may be used (e.g., Richardson & DiBenedetto, 1985; St. Louis & Ruscillo, 1987).

Thus, reliability evidence for criterion-referenced measures may sometimes take a form that is neither familiar nor straightforward, given expectations derived from years of confidently evaluating the reliability of norm-referenced measures. At least one expectation, however, can be shared for both types of measures—namely, that a wealth of evidence supporting high reliability counts for nothing without evidence supporting validity. Stated more specifically in terms relevant to criterion-referencing, a classification decision will mean nothing if it is consistent but incorrect.

(4) Careful Description of the Test Takers Used in Studies of Reliability and Validity

“The composition of the validation sample should be described in as much detail as is practicable. Available data on selective factors that might reasonably be expected to influence validity should be described.” (AERA, APA, & NCME, 1985, p. 14). “The procedures that are used to obtain samples of individuals, groups, or observations for the purpose of estimating reliabilities and standard errors of measurement, as well as the nature of population should be described” (AERA, APA, & NCME, 1985, p. 20).

Although the content of this guideline was touched on under the heading of validity above, isolating it as a separate guideline is meant to further emphasize the specificity required in clinicians’ judgments of test adequacy. The relevance of this guideline to criterion-referenced measures might easily be overlooked because score interpretation does not hinge on a comparison of the child’s score to the scores of a specified group, as it does for norm-referenced measures. Nonetheless, an appreciation of ways in which the child to be tested may differ from those on whom the test was developed can inform judgments regarding the appropriateness of the measure.
Further, such information can enrich the inferences to be drawn should the test user decide that the test sample is sufficiently similar to the child to warrant its use.

Let’s return to the example used to discuss implementation of the validity guideline—a language scale to be used with a child who speaks a dialect other than that for which the measure was initially designed. If information provided by the test author suggests that the measure has been studied using appropriately large groups of children, including those from the child’s minority group, then relatively straightforward use of the measure for description of the child’s language skills can follow. Alternatively, if it appears that the test had not been studied for such children, the clinician may find no more acceptable alternative measures and so may decide to use the instrument, but include appropriate cautionary statements in its interpretation and supplement its use with an informal measure designed to explore apparent strengths and weaknesses suggested by the standardized measure. Without information provided by the test author concerning the sample used in validation studies, however, the test user is not forewarned of the need for caution.

This guideline is, of course, quite relevant to both norm-referenced and criterion-referenced measures, but may be much more commonly followed for norm-referenced measures because of the centrality of comparisons between the test taker and the groups studied during the test’s development.

(5) Detailed Description of Test Administration

“The directions for test administration should be presented with sufficient clarity and emphasis so that it is possible to approximate for others the administrative conditions under which the norms and the data on reliability and validity were obtained” (AERA, APA, & NCME, 1985, p. 29).

The reference to norms in the quotation given above will be irrelevant for criterion-referenced measures that are not also norm-referenced. Otherwise, however, this recommendation is vitally important for the appropriate use of standardized criterion-referenced measures. Knowledge of and compliance with recommended procedures during test administration increase the likelihood that the measure will function as intended.

(6) Description of Test User Qualifications

“Test manuals should identify any special qualifications that are required to administer a test and to interpret it properly. Statements of user qualifications should identify the specific training, certification, or experience needed” (AERA, APA, & NCME, 1985, p. 36).

The importance of this guideline increases as the complexity of the response required from the test taker or the scoring and test interpretation processes used by the clinician increases. The availability of such information allows potential test users to determine whether they are currently qualified to use the instrument and what actions (e.g., training, referral) will be needed if they are not. Thus, adequate description of the user’s qualifications will contribute to a measure’s appropriate use.

As an example, one might consider the use of the Stuttering Severity Instrument for Children and Adults (Riley, 1994) to help describe a specific child’s stuttering. Knowledge of the specific (e.g., experience with the specific instrument) and general qualifications expected of examiners (e.g., previous experience assessing children with disfluencies) will help ensure that test users do not fail to recognize disfluencies and thereby underestimate the severity of the child’s problem because of their own inexperience.

As had been the case for the previous guidelines provided here, this guideline applies both to norm-referenced and criterion-referenced measures. In the case of norm-referenced measures, knowledge of expected test user qualifications helps preserve the integrity of the normative comparison; whereas for criterion-referenced measures, it preserves the integrity of the comparison to a performance criterion.

In summary, the six guidelines offered for the evaluation and selection of standardized criterion-referenced measures are closely related to those likely to be offered for similar purposes in relation to standardized norm-referenced instruments. With regard to some guidelines (e.g., that related to the need for clearly defined administration procedures), only small differences exist in the kinds of information being sought for the two kinds of measures; with regard to others (e.g., that related to reliability), an abundance of quite different, alternative methods exist for the newer, criterion-referenced measures, yet no clear front-runners have emerged to serve as the basis for explicit recommendations.

SUGGESTIONS FOR THE DEVELOPMENT AND USE OF INFORMAL CRITERION-REFERENCED INSTRUMENTS

In preceding sections, little of the proffered guidance relates to the informal criterion-referenced measures needed to answer many of the types of questions posed in clinical practice. Despite the need for such measures to be consistent (i.e., reliable) and useful for the purposes to which they are put (i.e., valid), they are rarely discussed in terms of reliability and validity. The Standards (AERA, APA, & NCME, 1985) indicate that although the guidelines can be “usefully applied to the full range of assessment techniques” (p. 4), less rigor can realistically be expected for measures that can be characterized as informal.

Recommendations regarding the appropriate use of informal measures are most often encountered in texts dealing with specific communication disorders (e.g., Bernthal & Bankson, 1993) and in research articles in which a set of tasks or items is described in considerable detail along with specific theoretical justifications for its use (e.g., Lewis & Freebairn, 1992). Such locations are desirable because they promote combined discussions of
what to measure and how to measure it within larger discussions of the nature of the disorder or how its damaging effects may be ameliorated or circumvented. Rarely, however, are informal criterion-referenced measures discussed within frameworks encouraging explicit attention to the promotion of their reliability and validity, with a few notable exceptions (e.g., Cole, Mills, & Dale, 1989; Leonard, Prutting, Perozzi & Berkley, 1978; Vetter, 1988). Vetter (1988) describes a process to be followed when the clinician has determined that alternative standardized measures are inappropriate or absent. In Figure 1, that process is illustrated using an example from George, the 7-year-old boy described at the beginning of this article. Whereas clinicians may wish to adopt this process fully only for those informal measures they are likely to use repeatedly, it articulates steps in decision making that can benefit the development of any informal measure.

Five major steps are concerned with the design of the measure:

1. identification of the specific question to be answered concerning the client;
2. selection of stimulus items that cover the desired content, are relevant to that content, and are of appropriate difficulty for the client;
3. simultaneous identification of expected, desirable responses that can reasonably be executed by the client and reliably scored by the clinician;
4. formulation of instructions likely to be understood by the client; and
5. development of decision-making guidelines, including performance guidelines.

In the case of George, the clinical question concerned his use of the grammatical morpheme *ing*. A set of stimuli was chosen that consisted of 20 pictures illustrating actions that could be described using the targeted form in combination with a variety of verbs likely to be in George’s vocabulary. Sentences were selected as the desired response because the clinician believed that George was capable of producing sentences of that complexity. The specific instruction that was used—“What’s this person doing?”—was designed to establish the need for the *ing* form. Sixteen correct items were selected as the criterion for passing on the somewhat arbitrary reasoning that 80% correct production would suggest a relatively well-established form in George’s production.

As illustrated in Figure 1, once the measure is administered and interpreted according to the predetermined design, additional steps depend on the measure’s success. If the measure failed to serve its intended purpose, the clinician would alter some aspect of the procedure’s design and use it in its revised form. On the other hand, if it appeared to perform as desired, the clinician would record its essential features for later use. In the case of George, George might produce a reasonable description of the picture but fail to use the targeted form, thus requiring either a modification of task instructions or the addition of a small number of training items. Alternatively, if implementation of the procedure were to suggest that George was able to follow


the instructions, but performed poorly in his use of *ing*, the clinician could incorporate that morpheme in treatment goals and save materials used for this probe for later use with other children, or with George after treatment had been undertaken using other materials.

Although the techniques proposed by Vetter (1988) are arguably cumbersome for fast-paced clinical lives, they reflect a method for thinking about criterion-referenced measures that addresses many basic, psychometric concerns. Thus, they can serve not only as recommendations for procedures to be used in the development of informal measures, but also as a summary for the guidelines provided in the earlier section of this article devoted to standardized criterion-referenced measures.
CONCLUSIONS

It is hoped that readers will conclude their reading of this article with a sense of having become somewhat more familiar with criterion-referenced measures—both by realizing what they already knew about such measures through use and familiarity with basic concepts underlying test use and by being exposed to discussions of their unique characteristics, use, and development. Repeated comparisons of norm- and criterion-referencing were intended to emphasize the intrinsic similarities of these approaches to score interpretation while acknowledging the incompatibilities that make their simultaneous application unwieldy, though not impossible.

Although the major thrust of this article was an educational one, at least one clinical need of a noneducational variety suggests itself. The relatively small number of standardized measures among those included in Table 1 was intended to reflect their relative scarcity. Indeed, standardized measures are lacking for many applications. Although deficiencies persist in the content knowledge required to support the development of such measures for many applications, continuing developments in the study of normal and abnormal development of communication skills (e.g., Smit, 1993; Windsor, 1994) are paving the way for future development of appropriate criterion-referenced measures in many areas.

In addition, a recent trend in educational and psychological testing (e.g., Bennett, 1993; Gitomer, 1993) involves a turning away from a dominant focus on the paper and pencil tests that have for so long provided a major focus for psychometricians in favor of the constructed response, or performance, measures more typical of communication measures. This trend may do much to supplement the technical knowledge needed for improvements to take place in our description of what children with speech-language deficits do and do not do in their efforts to communicate.

ACKNOWLEDGMENTS

The author thanks Chris Murphy, Marty Houghton, and Elena Plante, as well as several anonymous reviewers, for their help with this manuscript. A somewhat belated thanks is also extended to Ronald Berk and Darrell Sabers, who some time ago introduced me to the concept of criterion-referencing. Sabers' taped lectures, in particular, provided me with desert travel time that was far more enjoyable than even that provided by country music or talk radio.

REFERENCES


Received April 21, 1994
Accepted November 8, 1994

Contact author: Rebecca J. McCauley, Department of Communication Science, University of Vermont, Allen House, 461 Main Street, Burlington, VT 05405-0010.