Journal of Speech and Hearing Disorders, Volume 49, 34–42, February 1984

PSYCHOMETRIC REVIEW OF LANGUAGE AND ARTICULATION TESTS FOR PRESCHOOL CHILDREN

REBECCA J. McCAULEY LINDA SWISHER
The University of Arizona

Thirty language and articulation tests developed for use with preschool children were reviewed using ten psychometric criteria appropriate to norm-referenced tests. Half of the reviewed tests met no more than two criteria, and only three tests met over four criteria. Most frequently unmet criteria were those requiring empirical evidence of validity and reliability. Implications are drawn regarding the current status of norm-referenced language and articulation tests for preschool children.

Three assessment objectives are to determine the existence of a problem, to determine the goals of intervention, and to plan procedures for intervention. Despite recent attacks on the appropriateness of norm-referenced tests for any assessment objective (Muma, 1981; Muma, Lubinski, & Pierce, 1982; Muma & Muma, 1979), such tests are generally considered appropriate and necessary for the accomplishment of the first of these objectives—the identification of a speech or language impairment (Bloom & Lahey, 1978, p. 331; Launer & Lahey, 1981; Schery, 1981; Stark, Tallal, & Mellits, 1982).

Norm-referenced standardized tests help clinicians determine whether a child's score is like or unlike those of a group of individuals believed by the clinician to possess similar characteristics (Aram & Nation, 1982, p. 213; Salvia & Ysseldyke, 1981, p. 29; Schery, 1981). When a child's performance falls outside the range of scores received by most individuals in the normative sample, the child may be said to show significantly different and possibly nonnormal speech or language behavior.

The standardization and objective evaluation process required of norm-referenced tests gives the test user some confidence that a score received by the test taker reflects that person's actual level of skill and that it is not due to extraneous factors such as the way in which the test was given, the wording of instructions or test questions, or the setting in which the test was administered. Knowledge that a test possesses desirable psychometric characteristics offers an assurance to the test user that the effects of such extraneous factors have been reduced and that the behaviors assessed by the test are likely to be those of interest to the test user. In other words, this knowledge offers the potential test user some assurance that the test can help in the identification of impairment.

Several reviews of the psychometric characteristics of speech and language tests are available (Buros, 1972, 1978; Darley, 1979; Kilburg, 1982; Sommers, Erdige, & Peterson, 1978; Weiner & Hoock, 1973). These reviews are limited in their usefulness either because they fail to apply a single set of criteria to all tests or because they fail to consider more than a few tests. Reviews appearing in edited collections (e.g., Buros, 1972, 1978; Darley, 1979) differ somewhat in emphasis from one review to the next because they were written by different authors who did not apply identical criteria. Reviews in which a single set of criteria has been applied (e.g., Kilburg, 1982; Sommers et al., 1978; Weiner & Hoock, 1973) address only a small number of tests. This paper, therefore, applies the same criteria to 30 language and articulation tests developed for use with preschool children.

The intent of this paper is to stimulate discussion of the psychometric characteristics of language and articulation tests rather than to serve as a definitive psychometric review. Therefore, the criteria used in this review are a selected sample of a larger number of important psychometric criteria. Adherence to more numerous and, in some cases, stricter guidelines is commonly considered necessary for a well-developed norm-referenced test [American Psychological Association (APA), 1974].

Prior to a discussion of the specific criteria used in the review reported here, three concepts that are important to an understanding of norm-referenced tests and the standardization process will be discussed briefly. These are (a) test validity and reliability, (b) the normative sample, and (c) test norms and derived scores.

BACKGROUND ON SOME BASIC PSYCHOMETRIC CONCEPTS

Validity and Reliability

A measurement instrument, in this case a language or articulation test, has psychometric validity if it measures what it is designed to measure and does so accurately. Thus, a ruler is a valid measurement instrument of length to the degree that it accurately measures length. Likewise, a test of receptive vocabulary is a valid measurement instrument of receptive vocabulary to the degree that it accurately measures receptive vocabulary. If performance on a vocabulary test is influenced by visual perception, reading ability, or other variables, its validity for its intended purpose is compromised.

Validity, then, is not an absolute quality possessed by a

0022-4677/84/4901-0034$01.00/0

measurement instrument independent of its function; an instrument that is quite valid for the measurement of one attribute may be a much less valid instrument for the measurement of something else (APA, 1974, p. 31; Messick, 1980). For example, a valid test of receptive vocabulary is not necessarily a valid instrument for the measurement of receptive language as a whole. If it is a valid instrument for that purpose, that fact is a coincidence, not something that automatically results from the test's standardization.

Test validity is largely the product of successful test construction and evaluation. For example, during the selection of items to be included in a test, methods such as item analysis are used to make sure that each item reflects the class of behaviors being assessed and adds to the variety of such behaviors represented in the test. Thus, item analysis contributes to the validity of a test in the way that assembly-line inspection of individual engine parts contributes to the final quality of an automobile engine.

Three kinds of validity have usually been considered important for any test that measures behavior and is used to make inferences about underlying abilities—construct, content, and criterion-related validity (APA, 1974, pp. 25–55). Although Messick (1980) has presented a compelling case for considering construct validity as the keystone of test development, these three kinds of validity have often been defined independently and evidence about them for a given test is weighed jointly.

Construct validity refers to the degree to which a test measures the theoretical construct it is intended to measure (Anastasi, 1976, p. 151). Construct validity is examined by a careful comparison of a test author's delineation of the construct to be tested to the test's actual content. Because the evaluation of construct validity is difficult and somewhat subjective, it was not conducted for tests in this review.

To assess the *content validity* of a test, an individual who possesses expertise on the behavior assessed by the test examines the scope of specific items and of all items taken together. Moreover, the way the target behaviors are measured is examined to see to what extent the test provides relevant information about the behavior being tested.

Face validity refers to the apparent content validity of a test that is examined superficially or by an untrained individual. Anastasi (1976, pp. 139–140) notes that face validity is an important attribute of a test because it may affect the test taker's and even the test user's attitudes towards the test. However, face validity is not a substitute for information regarding the content validity of a test.

The examination of the *criterion-related validity* of a test involves the collection of empirical evidence that scores on the test are related to some other measure of the behavior being assessed. That other measure is called the criterion measure. Typically, two kinds of criterion-related validity are considered important: concurrent and predictive validity.

The *concurrent validity* of a test is determined by assessing how closely an individual's test score is related to his or her score on a criterion variable that is measured at about the same time the test score is obtained. An example of a direct estimate of concurrent validity for a language test might entail comparing test scores with expert judgments of the severity of a child's language problem. To obtain an indirect estimate of the concurrent validity of a test, the test designer might compare the scores that a number of test takers received on the language test being studied with their scores on another, already validated, language test. If the two sets of scores are closely related, then they are assumed to be measuring the same thing. And because the criterion measure has already been validated, it is presumed that the "same thing" they are measuring is the target behavior.

The *predictive validity* of a test is examined by assessing how closely an individual's test score can be used to predict future performance on a criterion measure. For example, to estimate the predictive validity of an articulation test, the test designer might determine how well the score an individual receives on the test under scrutiny can predict that same individual's performance on another, validated articulation test several months later, with no therapy given during the interim. The information that would be supplied by a test validated in this fashion could provide an important basis for the identification of subgroups of clients—subgroups differentiated by prognosis.

Any characteristic of a test that causes it to measure something other than the behavior of interest decreases the validity of that test. One of the characteristics that can reduce validity is the length of a test; if it is too long, fatigue may adversely affect performance on later items. Another such characteristic is the "enabling" behaviors required of the test taker; if a child cannot read a verbal intelligence test, a valid score will not be obtained on it (Salvia & Ysseldyke, 1981, p. 111). Though not often given a great deal of attention for language tests, receptive language skills may be important enabling behaviors for most expressive language tests. That is, the child taking an expressive test may need to understand fairly elaborate oral instructions in order to respond appropriately to test items.

One of the most important characteristics of a test that limits its validity is its reliability; when a test is unreliable, it fails to measure consistently what it is intended to measure and, therefore, it is a less valid measurement instrument (Salvia & Ysseldyke, 1981, p. 110). *Reliability*, then, refers to the consistency with which a test measures a given attribute or behavior. A perfectly reliable measurement instrument consistently gives the same value when the same variable is being measured. For example, if a perfectly reliable ruler is used to measure the length of a particular object, repeated measurements will result in the same value each time. Likewise, if a perfectly reliable language test is used to measure language ability, an individual tested at different times during the same day will receive the same score each time. However, a high degree of reliability alone does not ensure validity. For example, if the so-called language test inadvertently measures shoe size rather than language ability, the

individual's scores would provide consistent, that is, reliable, but invalid measures of language ability.

The term *test-retest reliability* refers to the stability of test scores over time. For example, because language abilities are not thought to fluctuate very much from day to day, a reliable language test should not produce scores that change very much over short periods of time. To offer evidence that a test shows test-retest reliability as part of its development, the test designer tests a group of individuals and then, within a relatively short time, administers the test instrument a second time. If the two scores each individual receives are closely and directly related (i.e., show a high positive correlation), then the test designer has obtained evidence of the test's stability over time.

Another kind of reliability that may be of interest for speech and language tests is *interexaminer reliability.* Evidence of this kind of reliability allows the test user to gauge the degree to which different test scorers or test administrators might influence test results. The test designer obtains this kind of evidence by determining whether sets of scores obtained by an individual on one test are closely related to those obtained when the test is given to the same person by several different administrators or when the test is scored by several different scorers. Evidence of interexaminer reliability is primarily of interest for tests in which the tester must make observations or classify responses as part of administering the test and for tests in which the presence or demeanor of the tester might greatly affect some test takers' responses.

The Normative Sample

Test norms are a statistical summary of the scores received by a normative sample. Because they are the basis for comparing a tested child with age and language-experience peers, norms obtained from different normative groups may be needed to provide different kinds of information regarding the existence of a problem. An example may help to illustrate this point. Suppose that a speech-language clinician has been asked to evaluate the language of a bilingual child. Now, first, suppose that the clinician is interested in seeing whether the child's language ability in English is significantly "poorer" than other children who have had similar language experiences. In this case, it would be important to use a language test that provides norms for bilingual children of the same age and language background as the child being assessed.

Next, suppose that the clinician is interested in seeing whether the bilingual child's ability in English is significantly different from other children who might be the child's classmates in school. To answer this aspect of the assessment question, it would be important to use a language test that had norms for a group of English-speaking children of the same age as the child being tested, regardless of their specific language experience.

Ideally, test designers would publish norms for a wide enough variety of groups to permit the test user to answer all relevant assessment questions for all possible test takers. Unfortunately, a representative normative sample for even the most common assessment question—Is there an impairment?—rarely appears in the data collected by the designers of speech and language tests. Nonetheless, the test manual should provide enough information about the normative sample so that the test user knows where the comparison of the individual's score to the norms may be weak.

For psychological tests, information regarding the age, geographic residence, and socioeconomic status of individuals in the normative sample should be supplied (APA, 1974, p. 21). These variables are thought to be important in affecting a child's test performance and, therefore, warrant consideration when language or articulation test results are interpreted. The effect of age on test performance is obvious: A 2-year-old child will surely perform less well than a normative sample of 10-year-olds. Geographic residence provides indirect information about the possible role of a nonstandard American English dialect on the language and articulation of the normative sample. If a child's dialect differs from the standard dialect—the dialect in which norm-referenced tests are almost always written—his or her performance on such tests may be adversely affected. Information about the socioeconomic status of members of the standardization sample is considered pertinent primarily because of findings that lower socioeconomic status can be associated with poorer test performance on existing language tests (e.g., Arnold & Reed, 1976; Johnson, 1974).

It is also important for the test user to know how the standardization sample was chosen. Sometimes, a test designer will exclude individuals on the basis of disability or some indication of nonnormal language ability. Although such exclusion may seem to make sense when one wants to ask whether a child's performance is similar or dissimilar to the performance of "normal" or "normally speaking" children, this approach to choosing the normative sample presents some difficulties. In particular, it means that even the most deviant scores contained in the normative sample represent normal performance. Because a child receiving a score just below the lowest score received by the normative sample has received a score with an unknown probability of occurrence, it may or may not reflect nonnormal performance (Salvia & Ysseldyke, 1981, p. 121). Thus, it may be harder to tell just how different a score needs to be before it reflects the possible presence of a language impairment.

Test Norms and Derived Scores

Three kinds of derived scores are frequently used during the presentation and interpretation of test norms and test takers' scores. These are developmental or age-equivalent scores, percentiles, and standard scores. These scores differ in the degree to which they make use of information about the central tendency and variability of scores obtained by the normative sample. The advantages and limitations of each kind of derived score will be

discussed briefly. More complete explanations can be found in Anastasi (1976, chapter 4) and Salvia and Ysseldyke (1981, chapter 5).

Age-equivalent or developmental scores are often considered a useful method of presenting test norms for behaviors or abilities that change considerably with age during childhood and adolescence. They are widely used to present test results to the parents of young clients.

Despite the fact that age-equivalent scores can provide the test user with information about the test taker's performance in relation to other children of the same chronological age, they present several difficulties (Salvia & Ysseldyke, 1981, pp. 66–69). Because the way in which they are calculated does not take into account individual differences, they fail to help the test user appreciate the range of performances that might be expected from a group of normal test takers. Thus, developmental scores are often difficult to interpret and "lend themselves to gross misinterpretation" (APA, 1974, p. 23). Specifically, their use can lead to questionable inferences about what one can expect from children with similar developmental scores but different chronological ages and, therefore, different amounts of exposure to language. For example, one might wrongly assume that two children, both with a developmental score of 6 years, but one age 7 and one age 9, would exhibit similar error patterns and potential for improvement. There is no evidence, however, that one can reasonably make that inference.

Percentile ranks indicate the percentage of individuals in the normative group whose test scores fell below a given value. For example, if Mary receives a test score with the percentile rank of 52, it means that 52% of the normative group received test scores lower than Mary's and that 48% received scores higher than hers. Because percentiles indicate the position of a test taker's score relative to the scores of the normative sample, percentiles make it easy to compare the test taker with the normative group.

On the other hand, percentiles have some disadvantages (Anastasi, 1976, pp. 78–79). One disadvantage is that small differences in percentile ranks at the high or low end of the percentile scale (e.g., below 10 or above 90) often reflect very big differences in raw scores because most individuals receive scores similar to the average score. For example, a difference of 20 points in two individuals' raw scores may appear as a difference of only a few percentile ranks if the scores are considerably higher or lower than the average, or a difference of a large number of percentile ranks if the scores are close to the average. This property of percentile ranks makes it impossible to determine the actual size of the raw score difference that separates individuals who receive different percentile ranks.

Standard scores are considered by some to be the most satisfactory kind of derived score for a number of reasons (Anastasi, 1976, p. 80). Although they are slightly more difficult to understand and calculate than percentiles or age-equivalent scores, their calculation makes use of information about the average score and variability of scores obtained by the normative sample. Because of this, standard scores can be more flexibly used than the derived scores discussed up to this point. They can be used to estimate the position of a test taker's score relative to the scores obtained by the normative sample, to compare one's score on two different tests, and to compare one person's score to someone else's in a meaningful way. The most common standard score is the z score. Because they are calculated in a way that preserves characteristics of the original set of raw scores, standard scores such as z scores (unlike age-equivalent scores and percentiles) can be used just like raw scores in later computations.

Although z scores and other kinds of standard scores are in most respects the preferred derived score, test manuals should always include information about the central tendency and variability of the test scores received by the normative sample. This is because these indices are the most basic descriptors of that sample's performance on the test (APA, 1974, p. 22).

PSYCHOMETRIC REVIEW

Methods

Initially, a list of 58 language and articulation tests available for use in the assessment of children was compiled using a variety of sources. These included published discussions and lists of tests found in Bernthal and Bankson (1981, pp. 205–219), Bloom and Lahey (1978, pp. 354–356), Darley (1979), and Salvia and Ysseldyke (1981, pp. 387–429). Among the tests listed were the five most frequently used norm-referenced tests of child language (Muma, Pierce, & Muma, 1983).

Because manuals for 16 tests were unavailable to us, manuals for 42 tests were examined to determine their apropriateness for this review. Twelve of these tests were considered inappropriate because their test manuals indicated that they were not meant to be used with preschool children or that they were not meant to be used as norm-referenced instruments and, therefore, lacked normative guidelines of any kind. Thus, 30 language and articulation tests were included in the final review process. (See Appendix for a complete list of these tests.) Test manuals and all other information that came with the test were examined in the review process.

The criteria used in the review consist of a set of characteristics that should be considered every time a clinician chooses or makes use of a norm-referenced test to evaluate a child's speech and language status. They constitute very basic information that the test user should have about any norm-referenced test (APA, 1974, pp. 9–10; Salvia & Ysseldyke, 1981, p. 32). They were chosen because of their recognized importance and relevance to tests of language and articulation, and because they could be translated into relatively objective decision rules. The failure of a test to meet these criteria or to provide information on them has serious effects on the merits of a test—no matter how in harmony the content of a given test is with the test user's concept of the skill being tested.

In this section, the psychometric criteria are first stated

generally, in a manner that would apply to many kinds of norm-referenced psychological tests. Then, the specific rules used to pass a test on that criterion in this review are stated for each criterion, and the consequences of failure on the criterion are discussed. These specific rules are suggestions for the way in which other users might implement the criteria when choosing language and articulation tests for preschool children.

Criterion 1. The test manual should clearly define the standardization sample so that the test user can examine its appropriateness for a particular test taker (APA, 1974, pp. 20–21; Weiner & Hoock, 1973). To pass this criterion, a test needed to give three pieces of information considered important for speech and language testing: (a) the normative sample's geographic residence, (b) socioeconomic status, and (c) the "normalcy" of subjects in the sample, including the number of individuals excluded because they exhibited nonnormal language or nonnormal general development (Salvia & Ysseldyke, 1981, pp. 118, 120–121, 394–395).

Consequences if unmet. Without this information, the test user cannot tell whether the normative sample is representative of the test author's intended population and whether it is the population against which the test taker's performance should be compared. For example, the way in which the standardization sample was chosen may have resulted in the use of children with a very different language background from that of the test taker. In such a case, the comparison of the test taker to the norms would be of little value if the clinician were interested in the normalcy of the test taker's language, but would be of considerable value if the clinician were interested in whether the test taker's language differed from that of the normative group.

Criterion 2. For each subgroup examined during the standardization of the test, an adequate sample size should be used (APA, 1974, pp. 27–28, 37). In order to pass this criterion, a test needed to have subgroups with a sample size of 100 or more. This particular value is consistently referred to by authorities as the lower limit for adequate sample sizes (Salvia & Ysseldyke, 1981, p. 123; Weiner & Hoock, 1973).

Consequences if unmet. If a small sample size is used, the norms are likely to be less reliable, that is, less stable; therefore, use of a different group of children might have resulted in different norms. In addition, relatively rare individuals (e.g., language- or articulation-impaired children) might not have been included in the sample because of the small number of subjects. This possibility makes the interpretation of the scores of possibly impaired children difficult. The smaller the sample size, the greater are the problems (Salvia & Ysseldyke, 1981, p. 123; Weiner & Hoock, 1973).

Criterion 3. The reliability and validity of the test should be promoted through the use of systematic item analysis during item construction and selection (Anastasi, 1976, p. 198). To pass this criterion, the test manual needed to report evidence that quantitative methods were used to study and control item difficulty, item validity, or both.

Consequences if unmet. Although the consequences would be less important if all other criteria to be discussed here were met by a test, that is rarely the case. Few norm-referenced tests provide adequate evidence that the test measures accurately what it purports to measure. This criterion, then, serves as an alternative indication that the test may possess validity and reliability. According to Anastasi (1976, p. 198), "high reliability and validity can be built into a test in advance through item analysis."

Criterion 4. Measures of the central tendency and variability of test scores should be reported in the manual for relevant subgroups examined during the objective evaluation of the test (APA, 1974, p. 22). To pass this criterion, both the mean and standard deviation had to be given for the total raw scores of all relevant subgroups.

Consequences if unmet. The mean is the average score received by members of the normative subgroup. The standard deviation gives the test user an estimate of how much variation was shown by the scores received by subgroup members. Because these pieces of information can also serve as the basis for other ways of presenting the norms (e.g., *z* scores), their absence robs the test user of flexibility in the use of the test norms.

Criterion 5. Evidence of concurrent validity should be supplied in the test manual (APA, 1974, pp. 26–27). To pass this criterion, the test manual needed to provide empirical evidence that categorizations of children as normal or impaired obtained using the test agree closely with categorizations obtained by other methods that can be considered valid, for example, clinician judgments or scores on other validated tests.

Consequences if unmet. The absence of this kind of evidence calls into question a test's ability to help with assessment questions related to the existence of impairment. Because the reason for using a norm-referenced test of language or articulation is to enable the test user to compare a child's score against the scores of other children as an aid in the determination of normalcy, the failure of a test on this criterion should cause the test user to question the usefulness of that test.

Criterion 6. Evidence of predictive validity should be supplied in the test manual (APA, 1974, pp. 26–27). To pass this criterion, a test manual needed to include empirical evidence that it could be used to predict later performance on another, valid criterion of the speech or language behavior addressed by the test in question.

Consequences if unmet. Evidence of this kind of predictive validity enables the test user to use a test to make assessment decisions related to the need for therapy. Therefore, its absence means that other, possibly invalid sources of information will be weighed more heavily in the decision process.

Criterion 7. An estimate of test-retest reliability for relevant subgroups should be supplied in the test manual (APA, 1974, pp. 50, 54). To pass this criterion, the test manual needed to supply empirical evidence of test-retest reliability, including a correlation coefficient of .90 or better (Salvia & Ysseldyke, 1981, p. 98) that was statistically significant at or beyond the .05 level (Anas-

tasi, 1976, pp. 108–109).

Consequences if unmet. Without this information, the test user does not know to what extent test results are stable and to what extent they will fluctuate over time. If this kind of reliability is low, the test should be viewed with the suspicion one would have of a rubber ruler. A correlation coefficient of .90 is considered a minimum standard of reliability when a test is used to make decisions regarding the existence of a problem (Salvia & Ysseldyke, 1981, p. 98). If in addition to being relatively high the correlation coefficient is not also statistically significant, that fact may be due to the specific characteristics of the individuals chosen as part of the normative sample, but would not be duplicated if the standardization process were repeated with other, similar groups.

Criterion 8. Empirical evidence of interexaminer reliability should be given in the test manual (APA, 1974, p. 50). To pass this criterion, a test manual needed to report evidence of interexaminer reliability that included a correlation coefficient of .90 or better (Salvia & Ysseldyke, 1981, p. 98) that was statistically significant at or beyond the .05 level (Anastasi, 1976, pp. 108–109).

Consequences if unmet. Without an estimate of interexaminer reliability, the test user does not know the degree to which a test taker is likely to receive similar scores if the test is given again by different individuals or if the same test is scored by different individuals. Neither can it be known whether the tester is likely to affect the scores of test takers in a way that improperly penalizes or favors them.

Criterion 9. Test administration procedures should be described in sufficient detail to enable the test user to duplicate the administration and scoring procedures used during test standardization (APA, 1974, p. 18). To pass this criterion, the test manual needed to provide sufficient description so that, after reading the test manual, the reviewer believed she could administer and score the test without grave doubts about correct procedures.

Consequences if unmet. Unless the administration process is described in detail, the test user does not know whether it is reasonable to compare the test taker's performance to the norms. Information supplied in the test manual should include specific instructions on how to administer, score, and interpret the test and on the setting in which testing should be conducted. If a test is administered without duplicating the procedures followed during standardization, the test taker may be given an unfair advantage or may be unfairly penalized by differences in instructions, surroundings, and so forth.

Criterion 10. The test manual should supply information about the special qualifications required of the test administrator or scorer (APA, 1974, p. 15; Salvia & Ysseldyke, 1981, p. 18). To pass this criterion, the test manual needed to state both general and specialized training required for administrators and scorers.

Consequences if unmet. Information of this kind should be given for all tests because the administration, scoring, and interpretation of test results should only be done by a qualified person. Moreover, for rating scales this information is particularly crucial because such scales use observations as test data. Without this information, it is impossible for the test user to judge the quality of the data obtained with the test.

Results

The examination of each test was first carried out by one of the authors (R.M.). In order to ensure the accuracy of the test ratings, a second examiner then rated each test. This examiner was an undergraduate research assistant with little formal training in psychometric theory. Prior to rating the tests, she read the background and methods sections of this paper and discussed with the first examiner the specific criteria to be used. The second examiner's rating agreed with those of the first examiner on 84% of the 300 rating judgments (30 tests × 10 criteria). Percentage agreement ranged from 67% on item analysis (Criterion 3) to 100% on interexaminer reliability (Criterion 8). For each rating of pass or fail on which the examiners disagreed, the test manual was consulted and the disagreement was reconciled. The process resulted in the modification of the results obtained by the first examiner in 7 instances, that is, for 2% of the ratings.

In Table 1, the number of tests meeting each criterion is summarized for the final rating results. Results are recorded for each category of test (articulation, vocabulary, and other language tests) and for all 30 tests combined. The category *other language tests* was used to encompass tests that varied considerably in the modalities or domains they addressed. Norm-referenced tests of auditory discrimination, which might be considered tests of receptive phonology, were not included in the review because of their questionable value to speech-language pathologists in their current form (Locke, 1980; Rees, 1973).

Half of the 10 psychometric criteria were met by fewer than six tests or by less than 20% of those tests included in the final review. Frequently unmet criteria dealt with the description of the sample (Criterion 1), evidence of concurrent validity (Criterion 5), predictive validity (Criterion 6), test-retest reliability (Criterion 7), and interexaminer reliability (Criterion 8). The two criteria concerned with predictive validity (Criterion 6) and interexaminer reliability (Criterion 8) were not met by any of the reviewed tests, and one test met the criterion for test-retest reliability (Criterion 7).

Inspection of the percentages attained by the different test categories shows no outstanding differences between categories. The number of psychometric criteria met by tests is quite small in each category. Possible exceptions to the overall similarity of results across test categories are found when vocabulary tests are compared with the other two categories: Inspection of the entries in Table 1 suggests that a larger percentage of vocabulary tests made use of item analysis (Criterion 3), yet a smaller percentage of vocabulary tests provided an adequate description of the test procedures (Criterion 9).

Table 2 lists the names of tests meeting each criterion. Although certain names appear more often than others, the pattern of met criteria observed in Table 1 is not due to the existence of a few tests that met almost all criteria

TABLE 1. The number of tests meeting each of 10 psychometric criteria, organized by categories of test content. (Percentage of tests in each category meeting each criterion is given in parentheses.)

Test category	*1 Descrip. of normative sample*	*2 Sample size*	*3 Item analysis*	*4 $\bar{x}$ (SD)*	*5 Concurrent validity*	*6 Predictive validity*	*7 Test-retest reliability*	*8 Inter-examiner reliability*	*9 Descrip. of test procedures*	*10 Descrip. of examiner qualifications*
Articulation (n = 5)	0	0	1(20%)	2(40%)	0	0	0	0	4(80%)	0
Vocabulary (n = 6)	1(17%)	1(17%)	4(67%)	1(17%)	0	0	0	0	3(50%)	3(50%)
Other language tests (n = 19)	2(11%)	5(26%)	4(21%)	4(21%)	5(26%)	0	1(5%)	0	18(95%)	11(58%)
All tests	3(10%)	6(20%)	9(30%)	7(23%)	5(17%)	0	1(3%)	0	25(83%)	14(47%)

while most met none. This fact is illustrated graphically in Figure 1 which records the number of tests that met one or more criteria, two or more criteria, and so on. Eight was the largest number of criteria met by any test and only one test (Newcomer & Hammill, 1977, see Appendix) met that many criteria. Two tests met the next largest number of criteria—five. The median number of criteria met by all tests was two; thus, half of the reviewed tests met two or fewer criteria.

Discussion

The pattern of met criteria that emerges from this review is possibly discouraging, yet in some ways predictable. The two criteria that would require the smallest expenditure of time and financial resources were met most often—Criterion 9 (description of procedures) and Criterion 10 (statement of tester qualifications). On the other hand, those criteria that require the application of considerable psychometric expertise, time, and money—criteria related to empirical evidence of validity and reliability—were met least often.

Although it may be argued that the failure of tests to pass these criteria was due to the very specific form in which the criteria were implemented, the use of broader requirements probably would not have affected the results greatly; many test manuals made no mention whatsoever of topics such as item analysis or empirical evidence regarding validity or reliability.

TABLE 2. Tests meeting each criterion. Test name abbreviations are those given in Appendix.

Criterion	*Number of tests*	*Tests*
1. Description of normative sample	3	ITPA, PPVT-R, TOLD
2. Sample size	6	BTBC, ITPA, PPVT-R, STACL, SOLST, TOLD
3. Item analysis	9	BTBC, EOWPVT, ITPA, PPVT, PPVT-R, QT, SICD, T-D, TOLD
4. Means and standard deviations	7	ICLAT, PAT, PPVT, T-D, TACL, TOLD, TTC
5. Concurrent validity	5	CELI, PLAI, PLST, TOLD, VLDS
6. Predictive validity	0	–
7. Test-retest reliability	1	TOLD
8. Interexaminer reliability	0	–
9. Description of test procedures	25	ACLC, BLST, BTBC, CELI, DASE, EOWPVT, ICLAT, ITPA, LAS, NSST, PAT, PLAI, PPVT, PPVT-R, REEL-scale, STACL, SICD, SOLST, T-D, TACL, TOLD, TTC, UTLD, VANE-L, VLDS
10. Description of tester qualifications	14	BLST, ICLAT, ITPA, PLAI, PPVT, PPVT-R, QT, SICD, STACL, SOLST, TACL, TOLD, UTLD, VANE-L

SUMMARY AND CONCLUSIONS

A review of 30 language and articulation tests designed for use with preschool children suggests that such tests fail to exhibit many of the psychometric characteristics required of well-standardized norm-referenced tests. On the whole, the reviewed tests failed to provide compelling empirical evidence that they can reliably and validly be used to provide information concerning the existence

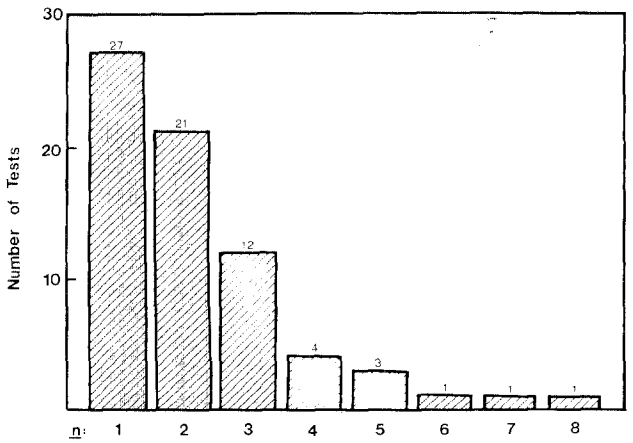

FIGURE 1. The number of the 30 reviewed tests that met at least n criteria.

of language or articulation impairment. These findings suggest important limitations on the use of such tests that must be considered by investigators and by speech-language clinicians.

Most failures of tests to meet individual criteria occurred as a result of an absence of sought-after information rather than as a result of reported poor performance on them. The tests were not shown to be either well developed or poorly developed. This fact may falsely comfort some readers who may assume that, if collected, the data on their favorite test would be favorable. However, when given no information about a psychometric characteristic, the test user is realistically left to wonder whether or not a test is invalid and unreliable for his or her purposes. Stated differently, no news is bad news.

At this point, test-using readers may rightly feel that they have been put in a most uncomfortable position. Indeed, we are all in the uncomfortable position of having less than perfect instruments available to help us compare to relevant norms the performance of children at risk for language or articulation impairment. Rather than responding to this sad state of affairs by discarding the idea that norm-referenced tests are potentially useful, test users have at least two constructive responses they can make. First, test users can make themselves more aware of psychometric principles and of the psychometric flaws of the tests they use or consider using, and can thereby reduce the impact of those flaws on clinical decisions. Clinical decisions are never properly based on test results alone; the speech-language clinician weighs test results in combination with other kinds of objective and subjective evidence when reaching a clinical decision. Where test results are severely undermined by flaws in construction, they should be weighed less heavily in the decision process.

Second, test users can realize the potential influence they wield in their role as test buyers by conducting a psychometric review prior to their test purchases. Competing tests might be evaluated in a manner similar to that used in this review. The outcome of this practice would be that test authors and publishers would be rewarded for the admittedly difficult and expensive steps involved in the production of psychometrically adequate tests. Test authors and publishers would then be encouraged to include the gathering of empirical evidence about reliability and validity as an integral part of preparing a test for publication, not as an expendable luxury.

ACKNOWLEDGMENTS

This research was conducted while R. McCauley was supported by National Institute of Neurological and Communicative Disorders and Stroke, National Research Service Award, 5 F32 NS06390 CMS. A portion of the investigation was presented at the Annual Convention of the American Speech-Language-Hearing Association in Toronto, 1982. The authors wish to acknowledge the helpful contributions of John Brockmann, Ralph Shelton, and the late Arlene Matkin. Special thanks are extended to Richard Curlee for his valuable comments and suggestions on an early version of this paper.

REFERENCES

American Psychological Association. (1974). *Standards for educational and psychological tests.* Washington, DC: APA.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.

Aram, D., & Nation, J. (1982). *Child language disorders.* St. Louis, MO: Mosby.

Arnold, K., & Reed, L. (1976). The Grammatic Closure subtest of the ITPA: A comparative study of black and white children. *Journal of Speech and Hearing Disorders, 41,* 477–485.

Bernthal, J., & Bankson, N. (1981). *Articulation disorders.* Englewood Cliffs, NJ: Prentice-Hall.

Bloom, L., & Lahey, M. (1978). *Language development and language disorders.* New York: Wiley.

Buros, O. (Ed.) (1972). *Seventh mental measurements yearbook.* Highland Park, NJ: Gryphon Press.

Buros, O. (Ed.) (1978). *Eighth mental measurements yearbook.* Highland Park, NJ: Gryphon Press.

Darley, F. (Ed.) (1979). *Evaluation of appraisal techniques in speech and language pathology.* Reading, MA: Addison-Wesley.

Johnson, D. (1974). The influences of social class and race on language test performance and spontaneous speech of preschool children. *Child Development, 45,* 517–521.

Kilburg, G. (1982). The assessment of emerging communication. *Communication Disorders, 7,* 87–101.

Launer, P., & Lahey, M. (1981). Passages: From the fifties to the eighties in language assessment. *Topics in Language Disorders, 1,* 11–30.

Locke, J. (1980). The inference of speech perception in the phonologically disordered child. Part 1: A rationale, some criteria, the conventional tests. *Journal of Speech and Hearing Disorders, 45,* 431–444.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Muma, J. (1981). *Language primer for the clinical fields.* Lubbock, TX: Natural Child Publishing.

Muma, J., Lubinski, R., & Pierce, S. (1982). A new era in language assessment: Data or evidence. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 7). New York: Academic Press.

Muma, J., & Muma, D. (1979). *MAP: Muma Assessment Program.* Lubbock, TX: Natural Child Publishing.

Muma, J., Pierce, S., & Muma, D. (1983). Language training in speech-language pathology: Substantive domains. *Asha, 25,* 35–40.

Rees, N. (1973). Auditory processing factors in language disorders: The view from Procrustes' bed. *Journal of Speech and Hearing Disorders, 38,* 304–315.

Salvia, J., & Ysseldyke, J. (1981). *Assessment in special and remedial education* (2nd ed.). Boston: Houghton Mifflin.

Schery, T. (1981). Selecting assessment strategies for language-disordered children. *Topics in Language Disorders, 1,* 59–73.

Sommers, R., Erdige, S., & Peterson, M. (1978). How valid are children's language tests? *Journal of Special Education, 12,* 393–407.

Stark, R., Tallal, P., & Mellits, E. (1982). Quantification of language abilities in children. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 7). New York: Academic Press.

Weiner, P., & Hoock, W. (1973). The standardization of tests: Criteria and criticisms. *Journal of Speech and Hearing Research, 16,* 616–626.

Received September 17, 1982
Accepted September 23, 1983

Request for reprints should be sent to Linda Swisher, Ph.D., Department of Speech and Hearing Sciences, University of Arizona, Tucson, AZ 85721.

APPENDIX

Thirty Language and Articulation Tests Examined During Test Review

Ammons, R., & Ammons, C. (1962). *The Quick Test* (QT). Provisional Manual. Psychological Reports Monograph Supplement. Supplement 1–7.

Ammons, R., & Ammons, H. (1948). *Full Range Picture Vocabulary Test* (FRPVT). Missoula, MT: Psychological Test Specialists.

Arlt, P. (1977). *Illinois Children's Language Assessment Test* (ICLAT). Danville, IL: Interstate Printers & Publishers.

Bankson, N. (1977). *Bankson Language Screening Test* (BLST). Baltimore: University Park.

Blank, M., Rose, S., & Berlin, L. (1978). *Preschool Language Assessment Instrument* (PLAI). New York: Grune & Stratton.

Boehm, A. (1971). *Boehm Test of Basic Concepts* (BTBC). Form A. New York: The Psychological Corporation.

Bzoch, K., & League, R. (1971). *The Receptive Expressive Emergent Language Scale* (REEL-scale). Gainesville, FL: Language Education Division, Computer Management Corporation.

Carrow, E. (1973). *Test for Auditory Comprehension of Language* (TACL). Austin, TX: Urban Research Group of Educational Concepts.

Carrow, E. (1974). *Carrow Elicited Language Inventory* (CELI). Boston: Teaching Resources Corporation.

Carrow, E. (1977). *Screening Test for Auditory Comprehension of Language* (STACL). Boston: Teaching Resources Corporation.

DiSimoni, F. (1978). *The Token Test for Children* (TTC). Hingham, MA: Teaching Resources Corporation.

Drumright, A. (1971). *The Denver Articulation Screening Examination* (DASE). Denver: Ladoca Project and Publishing Foundation.

Dunn, L. (1965). *Peabody Picture Vocabulary Test* (PPVT). Form A. Circle Pines, MN: American Guidance Services.

Dunn, L., & Dunn, L. (1982). *Peabody Picture Vocabulary Test-Revised* (PPVT-R) Form L. Circle Pines, MN: American Guidance Services.

Edmonston, W. (1963). *Laradon Articulation Scale* (LAS). Form Y. Beverly Hills: Western Psychological Services.

Foster, R., Giddan, J., & Stark, J. (1973). *Assessment of Children's Language Comprehension* (ACLC). Palo Alto: Consulting Psychologists Press.

Gardner, M. (1979). *Expressive One-Word Picture Vocabulary Test* (EOWPVT). East Aurora, NY: Slosson Educational Publications.

Hannah, E., & Gardner, J. (1974). *Preschool Language Screening Test* (PLST). Northridge, CA: Joyce Publications.

Hedrick, D., Prather, E., & Tobin, A. (1975). *Sequenced Inventory of Communication Development* (SICD). Seattle: University of Washington Press.

Hejna, R. (1959). *Developmental Articulation Test* (DAT). Ann Arbor, MI: Speech Materials.

Kirk, S., & Kirk, W. (1968). *The Illinois Test of Psycholinguistic Abilities* (ITPA) (Rev. ed.). Urbana, IL: University of Illinois Press.

Lee, L. (1971). *Northwestern Syntax Screening Test* (NSST). Evanston, IL: Northwestern University Press.

McDonald, E. (1968). *A Screening Deep Test of Articulation* (SDTA). Pittsburg: Stanwix House.

Mecham, M. (1958). *Verbal Language Development Scale* (VLDS). Circle Pines, MN: American Guidance Services.

Mecham, M., Joy, J., & Jones, J. (1967). *Utah Test of Language Development* (UTLD). Salt Lake City: Communication Research Associates.

Nation, J. (1972). Vocabulary usage test (VUT). *Journal of Psycholinguistic Research, 1,* 221–231.

Newcomer, P., & Hammill, D. (1977). *Test of Language Development* (TOLD). Austin, TX: Empiric Press.

Pendergast, K., Dickey, S., Selmar, J., & Soder, A. (1969). *Photo Articulation Test* (PAT). Danville, IL: Interstate Printers & Publishers.

Stephens, M. (1977). *Stephens Oral Language Screening Test* (SOLST). Peninsula, OH: Interim Publishers.

Templin, M., & Darley, F. (1969). *Templin-Darley Tests of Articulation* (T-D) (2nd ed.). Iowa City, IA: University of Iowa. Screening test only examined in review.

Vane, J. (1975). *Vane Evaluation of Language Scale* (VANE-L). Brandon, VT: Clinical Psychology Publishing.

Van Riper, C., & Erickson, R. (1975). *Predictive Screening Test of Articulation* (PSTA) (4th ed.). Kalamazoo, MI: Western Michigan University, Continuing Education Office.