LSHSS

Review Article

Psychometric Characteristics of Single-Word Tests of Children's Speech Sound Production

Peter Flipsen Jr.[a] and Diane A. Ogiela[b]

Purpose: Our understanding of test construction has improved since the now-classic review by McCauley and Swisher (1984). The current review article examines the psychometric characteristics of current single-word tests of speech sound production in an attempt to determine whether our tests have improved since then. It also provides a resource that clinicians may use to help them make test selection decisions for their particular client populations.
Method: Ten tests published since 1990 were reviewed to determine whether they met the 10 criteria set out by McCauley and Swisher (1984), as well as 7 additional criteria.
Results: All of the tests reviewed met at least 3 of McCauley and Swisher's (1984) original criteria, and 9 of 10 tests met at least 5 of them. Most of the tests met some of the additional criteria as well.
Conclusions: The state of the art for single-word tests of speech sound production in children appears to have improved in the last 30 years. There remains, however, room for improvement.

Up to 56% of the caseloads of practicing speech-language pathologists (SLPs) may include children with disorders of speech sound production (Mullen & Schooling, 2010). A survey by Skahan, Watson, and Lof (2007) indicated that 74% of clinicians *always* included norm-referenced single-word tests when assessing such disorders. Together with clinicians who indicated that they *sometimes* included such procedures, the value rises to 89%. Such tests are also frequently used to either include or exclude participants in both theoretical and applied research (e.g., Ertmer, 2010; Preston, Brick, & Landi, 2013; Torrington Eaton & Bernstein Ratner, 2013). Given their widespread use by both clinicians and researchers, their integrity clearly warrants scrutiny.

Evaluation of norm-referenced tests typically involves examining their psychometric characteristics (i.e., looking at how they were constructed). McCauley and Swisher (1984) conducted a psychometric review of the then-available norm-referenced language and articulation tests using criteria established for norm-referenced tests in *Standards for Educational and Psychological Tests* by the American Psychological Association (APA, 1974). On the basis of their review, McCauley and Swisher (1984) concluded that "the reviewed tests failed to provide compelling empirical evidence that they can reliably and validly be used to provide information concerning the existence of language or articulation impairment" (pp. 40–41). This represented a rather stinging indictment of the state of the art in testing at that time. It also appears to have served as somewhat of a wakeup call for test developers. Findings from more recent but smaller-scale reviews of language tests by Mikucki and Larrivee (2006) as well as Friberg (2010) suggest that noticeable improvements have been made. An unpublished study by Mathias (2010) applied the criteria of McCauley and Swisher (along with examining construct validity, see below) to nine more recent tests of speech sound production. Similar to the other more recent reviews, the conclusion by Mathias was that the tests of speech sound production had improved. Although such conclusions are heartening, our understanding of test construction has also improved in recent years. Thus, although the criteria of McCauley and Swisher remain valid, they may no longer be sufficient. That is, some additional issues may also warrant examination. In addition, given that McCauley and Swisher focused broadly on both speech and language tests, and given that the focus of the current report is on tests of speech sound production, some issues specific to those types of tests

[a]Pacific University, Forest Grove, OR
[b]Idaho State University, Meridian
Correspondence to Peter Flipsen Jr.: flipsen@pacificu.edu
Editor: Marilyn Nippold
Associate Editor: Linda Larrivee
Received May 28, 2014
Revision received September 22, 2014
Accepted January 11, 2015
DOI: 10.1044/2015_LSHSS-14-0055

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

likely need to be considered. The following discussion attempts to integrate all of those issues.

Test Validity

McCauley and Swisher (1984) noted that one of the most important questions of interest when evaluating tests is whether they measure what they claim to measure. In other words, how valid are they? In addressing this question, McCauley and Swisher used the standards available at that time (APA, 1974). However, since then, the standards have been revised. According to its authors, since 1985 the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) have stopped focusing on specific types of validity, but rather emphasize that validity is "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" (p. 14). Thus, rather than types of validity, there are various types of *validity evidence* that can help test users determine if a test is valid for their purposes. The present authors recognize the importance of this distinction. However, in order to remain consistent with McCauley and Swisher and to compare psychometric soundness of current articulation tests to those available in 1984, the following discussion is organized according to the same validity criteria they used along with other types of validity and reliability evidence.

McCauley and Swisher (1984) discussed several forms of validity. The first of these is *construct validity*, or the degree to which a test maps onto the theoretical construct it is supposed to assess. However, as noted by McCauley and Swisher, "the evaluation of construct validity is difficult and somewhat subjective" (p. 35). Even limiting the discussion to tests of speech sound production, the difficulty and subjectivity remain; in particular there is a continuing debate about the nature of how speech is perceived, managed, stored, and generated by the brain and nervous system (see, e.g., reviews in Ball & Kent, 1997, and Baker, Croot, McLeod, & Paul, 2001). As such, and similar to McCauley and Swisher, construct validity is not specifically examined in the current review.

Another form of validity is *content validity*, or the degree to which a test measures the relevant behavior. In the current context, the question would be whether the test appears to allow for a valid examination of speech sound production. Although all of the tests considered herein appear to do that (i.e., they have what some call *face validity*), there has been a long-standing discussion about whether they evoke a fully representative sample of typical speaking performance (e.g., H. B. Klein & Liu-Shea, 2009; Morrison & Shriberg, 1992). The usual negative argument is that they do not do so because the task involved is limited to single-word productions, whereas most utterances involve more than single words. Defenders of such instruments, however, counter that although that may be true, compared to conversational speech samples, these tests (a) yield a consistent sample across speakers, (b) allow for generation of norm-referenced scores to help make eligibility decisions, and (c) guarantee a sample of all of the relevant phonemes. One way to improve the representativeness of the sample might be by sampling the phonemes in relative proportion to their frequency of use in the language; however, such an approach might result in a very large sample set that is likely impractical. As an alternative, the sounds might be weighted during scoring relative to their frequency of occurrence in the language. The author of one test considered in the current review (Fudala, 2000) took such an approach. Doing so raises at least two concerns, however. First, it is not clear whether frequency data based on adult usage (as used by Fudala, 2000) are appropriate for a test geared to children's speech. Although the relative frequencies in children's versus adults' speech do not appear to have been statistically tested, a brief examination of data compiled from earlier studies in Shriberg and Kwiatkowski (1982) suggests that such differences may exist. Second, any such adjustments would not change the fact that the task itself (i.e., single-word productions) may not represent the cognitive and motor demands of typical speaking situations.

A related issue (that also reflects content validity) is the nature of the vocabulary selected to evoke the speech sample. Single-word tests of speech sound production typically use picture stimuli representing common words. The assumption is that children will easily recognize the pictures and be able to recall the appropriate words spontaneously. In other words, it is assumed that performance will not be constrained by limited vocabulary skills. Such an assumption may be problematic, however, because of known comorbidities between speech sound disorders and expressive or receptive language disorders (Shriberg & Austin, 1998). Some older studies have examined this question more directly with findings showing differences in children's abilities to spontaneously name the pictures across tests, at different ages, and depending on normal versus not normal status (Eveleigh & Warr-Leeper, 1983; Madison, Kolbeck, & Walker, 1982; Shanks, Sharpe, & Jackson, 1970). In order to minimize the risk of the vocabulary influencing performance, some type of systematic item analysis would seem to be necessary. McCauley and Swisher (1984) reported that only one of the five articulation tests they examined included such an analysis. An item analysis may consist of a statistical analysis during test development, or it may involve some sort of systematic field testing of the items to evaluate children's real-world responses.

One aspect of content validity not considered by McCauley and Swisher (1984) was the standard of comparison. For example, should comparisons be limited to a Mainstream American English (MAE) standard or could other dialect standards (in particular, nonstandard dialects such as African American English [AAE] or Appalachian English) be used? Such a consideration is not trivial. Cole and Taylor (1990) tested 10 children ages 5;11 (years;months) to 6;11 from Mississippi who spoke AAE with three then commonly used single-word articulation tests. When judged against MAE standards, the findings indicated that seven of 10, six of 10, and three of 10 children would be classified

as having a disorder on the three tests, respectively. However, when AAE was used as the standard of comparison, the number of children classified as having disorders dropped to zero of 10, two of 10, and one of 10. On the other hand, Washington and Craig (1992) concluded that dialect adjustments on one of those same tests (Arizona Articulation Proficiency Scale–Second Edition [AAPS-2]; Fudala & Reynolds, 1986) made no clinically significant difference for diagnosis of 28 children ages 4;6 to 5;3 who spoke AAE in the Detroit, Michigan, metropolitan area. Knowing when and if scoring adjustments are needed has clear implications for both overall caseload size and decisions about who specifically should receive services. Pearson, Velleman, Bryant, and Charko (2009) found that the phonetic inventories of the two dialects are not all that matter, but that speakers of MAE and AAE have different developmental milestones in speech sound development. They studied 854 AAE-speaking children across the United States and found that despite the overwhelming similarities in the individual speech sounds used by the two dialects, some phonological segments are learned earlier in AAE than in MAE, and others are learned earlier in MAE. Additionally, the dialects also have different patterns of phonotactic development. Such differences may have an impact on test scores and on decisions about whether or not children demonstrate a speech sound disorder at a particular age.

An aspect of content validity specific to tests of speech sound production is the type of analyses permitted with each test. Historically, these tests have focused almost exclusively on the accuracy of consonant sounds. The presumption has been that clinicians spend the great bulk of their time working on consonants and rarely need to work on vowels. In recent years, however, disorders of vowel production have been getting more attention (e.g., Ball & Gibbon, 2002; Bharadwaj & Assmann, 2013). Thus, the inclusion of some type of examination of vowels would broaden the utility of these tests. Similarly, there is the framework of the analysis. Tests in this area have tended to treat each sound independently by only examining accuracy of production across traditional word positions (initial, medial, and/or final). With the shift toward more linguistically oriented interventions (see E. S. Klein, 1996; Williams, McLeod, & McCauley, 2010), clinicians have become more interested in analysis of broader error patterns. One particular pattern framework that has gained widespread attention is that of natural phonology (Stampe, 1979) where broader error patterns are examined in terms of *phonological processes*. These represent simplification patterns exhibited by young, typically developing children (e.g., final consonant deletion, stopping) who readily move beyond such patterns as they mature. It also includes similar patterns that are retained for prolonged periods by older children with speech sound disorders. Like vowel analysis, the ability to formally examine error patterns may improve the utility of a test.

A third type of validity discussed by McCauley and Swisher (1984) is *concurrent validity*. They defined this as "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" (p. 38). This definition presents some degree of ambiguity, however. Clinician judgments represent overall categorical decisions (i.e., normal vs. disorder), and scores reflect numerical values that are often simply one data point among many within a comprehensive assessment. A preliminary review of the currently available tests suggested that many test developers assumed that this type of validity referred to test scores, and, thus, this was the interpretation taken herein. Taking this position also permitted a clearer distinction between concurrent validity and diagnostic accuracy (to be discussed below). This approach also did not affect our ability to evaluate long-term trends in test development because none of the tests reviewed by McCauley and Swisher (1984) provided concurrent validity findings.

The final form of validity mentioned by McCauley and Swisher (1984) was whether the tests provided data on *predictive validity*. McCauley and Swisher defined predictive validity as the presence of "empirical evidence that could be used to predict later performance on another, valid criterion of the speech and language behavior addressed by the test in question" (p. 38). Given the broad scope of how humans use speech and language, many types of predictive studies would be possible. For example, does performance on a test of speech production predict speech production accuracy in adolescence or adulthood? Does it predict later reading skills, overall academic achievement, or adult occupational outcomes? None of the tests reviewed by McCauley and Swisher provided such data.

One additional aspect of test validity that was not addressed by McCauley and Swisher (1984) is *diagnostic accuracy*, or how likely children are to receive an appropriate diagnosis (i.e., normal vs. disorder) with the test versus some other approach. This might be examined at least two ways. First, it might be examined at the group level (i.e., does the average score for a group of children otherwise diagnosed as having a speech sound disorder differ from the average score for a group of children otherwise diagnosed as normal?). Although such an approach may be informative, it is not necessarily helpful for clinicians who are more interested in the diagnosis of individual children. Thus, a second and perhaps more useful approach might be to report a combination of both *sensitivity,* or how well the test identifies children who actually have a disorder as having a disorder, and *specificity,* or how accurately the test excludes those with typical development from the disorder diagnosis (Perona, Plante, & Vance, 2005).

Test Reliability

In addition to validity, McCauley and Swisher (1984) noted the need to consider how consistently these tests measure the behavior of interest (i.e., their reliability). As with validity, there are several possible forms of reliability. Two were specifically examined by McCauley and Swisher. First, there is the question of *test–retest reliability*, or how

much test scores might vary for the same individual from one test administration to the next by the same examiner. This taps into whether things such as time of day, attention, fatigue levels, mood, and so forth affect the test results. Perhaps equally important is *interexaminer reliability*, or how much test scores might vary depending on who is administering the test. Does examiner style and/or rapport developed between the child and the examiner matter enough to affect scores? Both of these types of reliability are commonly measured and reported using correlation coefficients. McCauley and Swisher applied a specific criterion for each of these two types of reliability with coefficients of at least .90 (significant at $p < .05$) as the minimum standard. For both types of reliability, McCauley and Swisher reported that none of the five articulation tests they examined met this criterion.

Two other characteristics of test construction related to reliability are relevant here and were also examined by McCauley and Swisher (1984). First, have the test developers provided sufficiently detailed instructions for test administration? Lacking such instructions, it is difficult to imagine different examiners (or even the same examiner) being consistent in administering the test. In particular, if the test was administered quite differently from how it was administered during generation of the norms, would that affect the scores obtained? Four of the five articulation tests examined by McCauley and Swisher were judged to have provided sufficient information in this area. The second characteristic related to consistency or reliability is the qualifications of the examiner. Can anyone administer it, or does it require special training or specific professional expertise and training? Do examiners need experience with test administration in general or with tests of speech production in particular? Do they need training and experience with scoring these specific kinds of tests? Is experience with identifying whether or not the correct target sound was produced sufficient, or (in the case of tests that report phonological patterns) is training with the application of such patterns required? Related to this is the question of whether or not examiners need (possibly supervised) experience with the specific test in question before being permitted to independently administer it. None of the five tests examined by McCauley and Swisher provided sufficient information about training and qualifications.

One characteristic related to reliability that was not examined by McCauley and Swisher (1984) is whether the test developers provide data on standard error of measurement (*SEM*; Hutchinson, 1996). Every test score is only a sample of the individual's ability or behavior. As such, every score is subject to measurement error that reflects variations in such things as the setting, the particular examiner, time of day, attention, and fatigue levels. Thus, it is related to both test–retest and interexaminer reliability. The *SEM* specifically represents the standard deviation that would be obtained if an average person took the test many times. In practical terms, an *SEM* provides an estimate of the margin of error (or confidence interval) around any particular score. In other words, it provides the possible range of scores within which the true score actually lies. This is important because it provides clinicians with some room for clinical judgment. For example, if a child achieves a standard score of 80 but the score on that test has an *SEM* of 5, the true score lies between 75 and 85. If the critical cutoff for eligibility for services is a score of 77, the clinician has some flexibility. If performance on other measures such as intelligibility or speech sound accuracy in conversational speech is poor, the clinician may infer that the lower end of the range is more representative of the child's abilities and recommend services. On the other hand, if performance on those other measures is more age appropriate, services might not be recommended. Likewise, if the same child achieved a score of 75 on a test of a related ability (e.g., expressive vocabulary) with an *SEM* of 7, the true score for that second test would be somewhere between 68 and 82. Because the ranges of the scores for the two tests overlap, it could be argued that the child demonstrated similar levels of ability in both areas. This would be particularly helpful when trying to identify general areas of relative strength and weakness. Another important aspect of *SEM* is that it can also serve as a metric of the precision of individual test scores on a given test (Harvill, 1991). If a test has a large *SEM*, it means that there would be greater variability in an individual's test performance, whereas a smaller *SEM* indicates less variability. Thus, a test with an *SEM* of ± 3 is more precise than a test with an *SEM* of ± 7. The larger an *SEM* is, the wider the confidence interval is around a score. If a confidence interval around a given score is very wide, then the information we can take from that score is fairly limited because it lacks precision.

Test Norms

A third relevant aspect of test psychometrics mentioned by McCauley and Swisher (1984) is the nature of the comparison sample (i.e., the norms). Of particular importance is whether the sample is sufficiently inclusive. Given a choice among tests, both clinicians and researchers want to know whether the children they assess are similar enough to the normative sample for particular tests to be appropriate. Therefore, a detailed a description of that normative sample is needed to make those determinations. Characteristics such as age range, geographic and ethnic distribution, gender split, and socioeconomic status (SES) are usually considered. McCauley and Swisher reported that none of the five articulation tests they evaluated provided sufficient information in this regard.

An issue that has been identified since the publication of the review by McCauley and Swisher (1984) is the profile of the individuals who are included in the normative sample group. There is specifically the important question of whether or not the sample should only include those with "normal" skills. This is what McFadden (1996) referred to as a *truncated sample*. Proponents of a truncated sample such as Peña, Spaulding, and Plante (2006) argue (among other things) that this makes identification of a disorder less ambiguous. The alternative would be that test developers

should simply "include individuals who represent the age and demographic characteristics of those for whom the test is intended" (Peña et al., 2006, p. 247). This latter sample would then also include a representative portion of those with the disorder in question. These are what McFadden called *full-range samples*, which she argued would (among other things) reduce the risk of children at the low end of the normal range being misidentified as having a disorder. Resolution of the truncated versus full-range sample question is beyond the scope of the current review, but identifying the type of sample used may assist both clinicians and researchers with test selection.

A commonly discussed aspect of the normative sample is its size. Without question, the larger the sample the better because larger samples are more likely to be normally distributed and, thus, are more likely to lead to an appropriate diagnosis. Overall sample size, however, is likely less important than the size of the specific subgroup of children to which any particular case is being compared. Normative tables (from which scores are derived) for tests of children are typically divided into multiple subgroups on the basis of somewhat narrow age ranges. This is intended to capture the rapid rate of change that occurs in many domains during the developmental period. A test may have a total normative sample of 1,000 children, but if there are 20 age groups in the norms tables, there would only be about 50 children per subgroup that clinicians and researchers will be comparing each child against. A commonly used standard is that each comparison subgroup should include at least 100 children; using this criterion, McCauley and Swisher (1984) reported that none of five tests they examined had sufficient children per subgroup.

One additional aspect of the norms included in the review by McCauley and Swisher (1984) was whether the test developers provided mean and standard deviation values for each of the subgroups. As they noted, "standard deviation gives the test user an estimate of how much variation was shown by the scores received by the subgroup members" (p. 49). They also suggested that it provides flexibility because it allows for generation of alternative scores such as *z* scores. Only two of five tests that they reviewed provided this information.

The final question is whether test developers should provide separate norms for boys and girls. Although not examined by McCauley and Swisher (1984), a recent review by McLeod (2013) indicated that some studies have noted gender differences in normal speech sound acquisition. When combined with consistent findings of a higher proportion of boys having speech sound disorders compared to girls (Bernthal, Bankson, & Flipsen, 2013, pp. 173–174), it suggests the need for test developers to at least examine the question.

The Current Review

The aim of this review article is to examine the currently available single-word tests of English speech sound production in children with regard to the psychometric characteristics described above. The intent is to (a) assess whether test developers are currently providing more psychometrically sound tests (i.e., have things improved since the review by McCauley & Swisher, 1984?), and (b) provide a catalog of how well each of these tests provide meaningful information relative to this expanded list of psychometric characteristics. This will assist clinicians and researchers with test selection and (if necessary) provide test developers with a basis for improving the diagnostic tools in this area.

Method

Tests Examined

The first author conducted a search of the Buros Center for Testing online test reviews and the American Speech-Language-Hearing Association online Directory of Speech-Language Pathology Assessment Instruments to identify norm-referenced, single-word tests of children's speech sound production. Only tests published since 1990 were considered in order to limit the review to tests likely to still be in regular use. Tests that focus mainly on speech motor skills, such as the Kaufman Speech Praxis Test for Children (Kaufman, 1995), or which involved secondary analysis of stimuli developed for other tests, such as the Khan–Lewis Phonological Analysis (Khan & Lewis, 2002), were excluded. The search yielded 10 tests, which are highlighted in Table 1.

As indicated in Table 1, all of the tests covered the age range from 3;0 to 7;11. Three (Arizona Articulation Proficiency Scale–Third Edition [AAPS-3; Fudala, 2000], Clinical Assessment of Articulation and Phonology–2 [CAAP-2; Secord & Donahue, 2014], and Goldman Fristoe Test of Articulation–Second Edition [GFTA-2; Goldman & Fristoe, 2000]) included even younger children, with one (AAPS-3) including children as young as 1;6. Most of the tests included children up to at least 8;11, with seven (AAPS-3; Bankson–Bernthal Test of Phonology [BBTOP; Bankson & Bernthal, 1990]; CAAP-2; GFTA-2; LinguiSystems Articulation Test [LAT; Bowers & Huisingh, 2010]; Smit–Hand Articulation and Phonology Evaluation [SHAPE; Smit & Hand, 1997]; and Structured Photographic Articulation Test II, Featuring Dudsbury [SPAT-D II; Dawson & Tattersall, 2001]) extending to even older ages. Three (AAPS-3, GFTA-2, LAT) could be used with adolescents, and the two extending the highest (GFTA-2, LAT) were designed to also include young adults (up to 21;11). The normative samples ranged in size from 650 to over 5,000 and included between 10 and 49 subgroups in the normative tables. The tests each included between 30 and 80 target words.

Evaluation Criteria

A total of 17 different criteria were examined, 10 of which had been included in the review by McCauley and Swisher (1984). Seven of the criteria were related to validity, with the first four being strict validity criteria. The first of these, which relates to content validity, was whether or not the stimuli had been selected using a formal item-analysis

Table 1. Tests reviewed.

Test name	Acronym	Reference	Norms age range	Total norms sample size	Norms age groups	No. of test words
Arizona Articulation Proficiency Scale–Third Edition	AAPS-3	Fudala, 2000	1;6–18;11	5,515	18	46
Bankson–Bernthal Test of Phonology	BBTOP	Bankson & Bernthal, 1990	3;0–9;11	1,070	22	80
Clinical Assessment of Articulation and Phonology–2	CAAP-2	Secord & Donahue, 2014	2;6–11;11	1,486	13	44
Diagnostic Evaluation of Articulation and Phonology (American Edition)	DEAP-A	Dodd, Hua, Crosbie, Holm, & Ozanne, 2006	3;0–8;11	650	11	30
Goldman Fristoe Test of Articulation–Second Edition	GFTA-2	Goldman & Fristoe, 2000	2;0–21;11	2,350	49	53
Hodson Assessment of Phonological Patterns–Third Edition	HAPP-3	Hodson, 2004	3;0–7;11	886	10	50
LinguiSystems Articulation Test	LAT	Bowers & Huisingh, 2010	3;0–21;11	3,030	25	52
Photo Articulation Test–Third Edition	PAT-3	Lippke, Dickey, Selmar, & Soder, 1997	3;0–8;11	800	20 (for boys) 14 (for girls)	72
Smit–Hand Articulation and Phonology Evaluation	SHAPE	Smit & Hand, 1997	3;0–9;0	2,091	10	80
Structured Photographic Articulation Test II, Featuring Dudsbury	SPAT-D II	Dawson & Tattersall, 2001	3;0–9;11	2,270	14	45

Note. Age is expressed in years;months.

procedure (either with some sort of statistical analysis and/or field testing process). Thus, it required more than simply asking for the opinions of clinicians about the appropriateness of the words. This helped ensure that most children would readily recognize and be able to spontaneously attempt a production of the words. The second validity criterion was whether the test developers reported formal evidence of concurrent validity. In this case, the present authors were looking for reports of how similar scores on a given test were to scores on another test in this same domain. The third validity criterion was whether the test developers reported findings on predictive validity; here manuals were examined for reports of whether test scores could predict future outcomes. To be consistent with McCauley and Swisher's original notion (see earlier definition), the net was cast broadly so that it might include any outcome for which competency with speech production might have an impact (e.g., speech skills, reading skills, academic performance, or occupational outcomes). The final validity criterion was diagnostic accuracy. This was examined by looking for reports of either (a) sensitivity and specificity or (b) group comparisons between those who had been defined as having a speech sound disorder by some other means (e.g., performance on another test or clinician opinion) and either a separate matched group of typically developing children or the test's normative tables.

The other three validity criteria did not neatly fit that broad category and thus were classified as *validity-related* criteria. The first was whether the test developers allowed for consideration of dialect variation. Did the test manual discuss the potential impact of dialect on test performance? If so, did it include specific guidance for using alternate scoring or making other adjustments? The second criterion was whether formal analysis of vowels was part of the test. Were vowels systematically included, and was performance on vowels incorporated into the test scoring? The third criterion involved pattern/process analysis. Did the test have a formal mechanism for describing errors using phonological process labels such as stopping, cluster reduction, and final consonant deletion? Given disagreements about what should or should not be included in the list of processes to be examined, the current authors chose to be neutral about this question. Any list of such processes was judged to be evidence that the test developers had included such an analysis.

Five criteria were related to reliability; the first two of these clearly involved reliability itself and included looking for evidence of test–retest and interexaminer reliability. In each case, the cutoff used was a report of correlation coefficients of at least .90 (similar to McCauley & Swisher, 1984). In cases where more than one coefficient was presented, all needed to be at least .90 to meet this criterion. The remaining reliability criteria were classified as *reliability-related*. The first of these involved asking whether the manuals provided clear administration instructions. Although somewhat subjective, the present authors looked for evidence that the test could be administered in a consistent manner using the instructions provided. Specific examiner qualifications were then evaluated. Did the test manual indicate specific training, academic preparation, or professional credentials were needed to administer the test? It should be noted that the present authors took a neutral position as to whether administering such tests should be limited to certified or licensed SLPs. The present authors also asked whether or not the test developers provided data on *SEM* (either as a table of *SEM* values or as confidence intervals for the scores).

Five criteria were ultimately examined with regard to the test norms. The first of these involved examining whether the normative sample was described in sufficient detail. Did it, for example, provide details about the geographic and ethnic distribution of the sample? Did it include considerations for SES? The second criterion involved determining whether the test developers chose to use either a

truncated or a full-range sample. Although it could be argued that this simply represents an aspect of the normative sample description, it was judged to be an important enough issue to be examined separately as well. Tests that failed to provide sufficient detail to evaluate this aspect of the sample were, of course, not considered to have been sufficiently described. Evidence in this case would be whether the manual stated that the normative sample specifically excluded or included individuals with speech sound disorders. The use of broad terms such as "receiving special education services" or "receiving speech and language services" that would normally include individuals with speech sound disorders was judged sufficient in this regard. The third norms criterion looked at the adequacy of the subgroup sample size. Were there at least 100 children in each of the subgroups presented in the normative tables? Then the present authors asked whether the test developers presented mean and standard deviation data for each of the subgroups. The fifth criterion asked whether the test developers considered gender differences. Did they present separate male and female data for use in scoring the test? If not, was the question of gender differences considered in developing the normative tables?

Evaluation Procedure

Each of the two authors independently reviewed the manuals for each of the tests under consideration to reach a preliminary conclusion about whether the tests met each of the psychometric criteria. Discussion then followed until consensus was reached for each criterion on each test.

Results

Validity

Findings for each of the tests relative to the validity and validity-related criteria are highlighted in Table 2. A formal process of item analysis was included in the development of five of 10 tests. Of these, the authors of the CAAP-2, GFTA-2, and the SPAT-D II did field testing of the items, and the authors of the Hodson Assessment of Phonological Patterns–Third Edition (HAPP-3; Hodson, 2004) and the Photo Articulation Test–Third Edition (PAT-3; Lippke, Dickey, Selmar, & Soder, 1997) conducted statistical analyses for internal consistency. It should be noted that the developers of the CAAP-2 only did field testing for the original version (not the second edition) of their test. Credit was given for meeting this criterion, however, because only one stimulus item (a new picture for the word *computer*) was changed between the two editions. Just over half (six of 10) of the tests (AAPS-3, CAAP-2, Diagnostic Evaluation of Articulation and Phonology–American Edition [DEAP-A; Dodd, Hua, Crosbie, Holm, & Ozanne, 2006], LAT, PAT-3, SPAT-D II) provided concurrent validity data. In each case, correlations were provided between performance by children on that test and performance on another single-word test. None of the tests presented data on predictive validity. Diagnostic accuracy data were available for six of 10 tests. Only two of the tests (CAAP-2, DEAP-A) included sensitivity and specificity findings (for the CAAP-2, this was presented as predictive validity findings; see below). Four other tests (HAPP-3, GFTA-2, PAT-3, SHAPE) included findings from groupwise comparisons either between typically developing children and children with speech sound disorders or showed that some subgroup of children with speech sound disorders performed significantly below the normative mean.

As to the validity-related criteria, dialect was the first to be addressed; the key question was whether or not the test developers discussed the issue of dialect and its potential influence on test scores. Almost half (four of 10) of the tests did not address the issue with regard to test administration or scoring at all. The six tests that did discuss dialect did so in a variety of ways and with varying degrees of detail (see the Discussion section). Formal vowel analysis was included in three of 10 tests (AAPS-3, DEAP-A, PAT-3), and a fourth test (SPAT-D II) allowed for vowel analysis, but vowel errors were not included in the test scoring. Phonological process/pattern analysis was possible with

Table 2. Validity and validity-related findings.

	Test acronym[a]									
Criteria	**AAPS-3**	**BBTOP**	**CAAP-2**	**DEAP-A**	**GFTA-2**	**HAPP-3**	**LAT**	**PAT-3**	**SHAPE**	**SPAT-D II**
Validity										
Formal item analysis conducted	No	No	Yes[b]	NR	Yes	Yes	No	Yes	No	Yes
Concurrent validity data	Yes	No	Yes	Yes	No	No	Yes	Yes	No	Yes
Predictive validity data	No	No	No	No	No	No	No	No	No	No
Diagnostic accuracy data	No	No	Yes[d]	Yes[d]	No	Yes[c]	Yes[c]	Yes[c]	Yes[c]	No
Validity-related										
Dialect differences discussed	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes
Formal vowel analysis	Yes	No	No	Yes	No	No	No	Yes	No	—
Phonological process analysis	No	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes

Note. Criteria in italics were included in McCauley and Swisher (1984). Yes = met criterion; No = failed to meet criterion; NR = not reported or unclear. Em dash indicates that this information was not included in score calculations.

[a]See Table 1. [b]Did item analysis on previous edition (only one modified test stimulus; see text). [c]Included data on group differences (normal vs. disorder) but did not report sensitivity/specificity. [d]Reported sensitivity and specificity data.

six of 10 tests (BBTOP, CAAP-2, DEAP-A, HAPP-3, SHAPE, SPAT-D II).

Reliability

As can be seen in Table 3, test–retest reliability findings were reported for eight of 10 tests, but of these, only three tests (AAPS-3, HAPP-3, SPAT-D II) met the criterion of all coefficients being at least .90. Four others (CAAP-2, DEAP-A, GFTA-2, LAT) presented mixed results consisting of multiple coefficients with values reported both above and below the .90 cutoff. One test manual (BBTOP) reported all values below .90. Relative to interexaminer reliability, findings were reported for seven of 10 tests. Four of the tests (CAAP-2, PAT-3, SHAPE, SPAT-D II) met the .90 criterion, and three others (DEAP-A, GFTA-2, HAPP-3) yielded mixed results. One additional test (AAPS-3) only reported findings for a much earlier version of the test. As such it was deemed to have not met the criterion.

For the reliability-related criteria, all 10 tests were judged to have provided clear administration instructions, and nine of 10 tests specifically stated examiner qualifications. The manual for the SHAPE mentioned SLPs several times but did not appear to explicitly state who can or cannot administer the test. *SEM* data were included in eight of 10 test manuals. The exceptions were the HAPP-3 and the SHAPE.

Test Norms

As can be seen in Table 3, relative to the norms, the developers of six of 10 tests provided descriptions of their normative samples that were judged to be sufficient for most applications; two tests (AAPS-3, SPAT-D II) failed here because they did not specify enough detail to know if the sample was truncated or full range (the next criterion). In addition, the manuals for three tests (BBTOP, PAT-3, SPAT-D II) did not include information about the SES of their normative group members. There was sufficient information available in eight of 10 test manuals to allow for classification of the normative samples as either truncated or full range. The exceptions were the AAPS-3 and the SPAT-D II. Six tests appeared to use full-range samples (CAAP-2, DEAP-A, GFTA-2, HAPP-3, LAT, PAT-3); one test manual (BBTOP) described their normative sample as including "a relatively large sample of normally developing children" (Bankson & Bernthal, 1990, p. 64). This suggested that children with speech sound disorders had not been included, and thus the sample was classified as truncated. Norms for the SHAPE involved a merger of two samples, one of which appeared to be truncated, whereas the other seemed to be full range. Subgroup size met the 100+ criterion for three tests (AAPS-3, LAT, SHAPE) but did not for another four tests (BBTOP, DEAP-A, GFTA-2, SPAT-D II). Subgroups in the remaining three tests (CAAP-2, HAPP-3, PAT-3) were mixed in size, with some subgroups not meeting the criterion. Data for means and standard deviations for the subgroups were presented for all 10 tests. Of the 10 tests that were examined, eight considered gender differences in scoring (the exceptions were BBTOP and CAAP-2). Three tests (DEAP-A, GFTA-2, PAT-3) provided norms for both boys and girls at all ages, and two (AAPS-3, SPAT-D II) provided separate norms

Table 3. Reliability, reliability-related, and norms findings.

	Test acronym[a]									
Criteria	**AAPS-3**	**BBTOP**	**CAAP-2**	**DEAP-A**	**GFTA-2**	**HAPP-3**	**LAT**	**PAT-3**	**SHAPE**	**SPAT-D II**
Reliability										
Test–retest reliability data[b]	Yes	No	[c]	[c]	[c]	Yes	[c]	NR	NR	Yes
Interexaminer reliability data[b]	NR[d]	NR	Yes	[c]	[c]	[c]	NR	Yes	Yes	Yes
Reliability-related										
Clear administration instructions	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Examiner qualifications stated	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	NR	Yes
SEM data included	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes
Norms										
Clearly defined norms sample	No[e]	No[f]	Yes	Yes	Yes	Yes	Yes	No[f]	Yes	No[e,f]
Truncated or full-range norms	NR	T	F	F	F	F	F	F	?[g]	NR
At least 100 per norms subgroup	Yes	No	[h]	No	No	[i]	Yes	[j]	Yes	No
Ms and SDs for subgroups	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Discussed gender differences	Yes[k]	No	No	Yes[l]	Yes[l]	Yes[m]	Yes[m]	Yes[l]	Yes[n]	Yes[o]

Note. Criteria in italics were included in McCauley and Swisher (1984). Yes = met criterion; No = failed to meet criterion; NR = not reported or unclear; *SEM* = standard error of measurement; T = truncated sample (i.e., norms sample did not include individuals' speech sound disorders); F = full-range sample (i.e., norms sample included individuals with speech sound disorders).

[a]See Table 1. [b]All correlation coefficient(s) at least .90. [c]Mixed findings (several coefficients reported but some <.90). [d]Only reported for earlier version of the test. [e]Did not discuss inclusion or exclusion of children with speech sound disorders. [f]Failed to include socioeconomic status data in norms information. [g]Two norms samples combined (one truncated, one full range). [h]100+ only for 10/13 subgroups. [i]100+ only for 6/10 subgroups. [j]100+ only for 15/20 subgroups. [k]Separate male/female (M/F) norms up to age 5;11 (years;months). [l]Separate M/F norms for all ages. [m]Discussed gender differences and pooled data (with specific justification). [n]No gender-based values for *z* scores but did have gender-based scores for use of phonological processes. [o]Separate M/F norms up to age 6;11.

for boys and girls at younger ages but then merged the groups at older ages. Two tests (HAPP-3, LAT) pooled all male and female data in the tables because the test developers concluded that there was little to no difference between them. One test, the SHAPE, did not have separate *z* scores by gender, but it did report gender differences in the description of phonological processes used at each age. The CAAP-2 manual mentioned gender differences relative to the normative population, but only to the extent of noting that the difference between the overall number of boys and girls (721 vs. 765) was minimal. No mention appears to have been made of performance differences by gender.

Discussion

The current review was set in the context of whether there has been any improvement in tests of speech sound production in children in the 30 years since the publication of the review by McCauley and Swisher (1984). An initial observation is that the number of available tests has doubled from five to 10 since 1984, thus providing clinicians with more options. That said, the focus herein was on whether or not the quality of the options has improved. Relative to the original 10 criteria set out by McCauley and Swisher, two tests (CAAP-2, SPAT-D II) met seven of them, and four others (AAPS-3, HAPP-3, LAT, PAT-3) met six of the 10. Three tests (DEAP-A, GFTA-2, SHAPE) met five criteria, and one test (BBTOP) met only three of the criteria. Compared to the findings of McCauley and Swisher in which none of the tests met even four of the criteria, this appeared to represent a significant improvement.

As to specific criteria, formal item analysis was confirmed for only five of the tests reviewed. One of the tests, the DEAP-A, reported a pilot study of the items that were replacing items in the British version of the test. However, the DEAP-A was rated NR (not reported or unclear) on this item because there was no mention of any sort of quantitative procedure (i.e., a statistical analysis) or systematic field testing of the items beyond the fact that the new items were presented to five children. Overall, the increase in the number of tests that conducted an item analysis does represent an improvement from McCauley and Swisher (1984), who reported that none of the five tests they reviewed had done so. However, the lack of such analyses for half of the tests raises a concern because clinicians can be less certain that children with comorbid expressive language impairments will be able to spontaneously name the test items. The same may be true even for some children without such challenges. In addition, the lack of such item analyses increases the risk that clinicians may have to base their evaluation on a mix of imitated and spontaneous productions; this raises questions about the validity of both the child's performance and the resulting diagnosis.

Six of the test manuals included findings for concurrent validity, allowing clinicians to see how children might perform on other articulation tests and related tests. In comparison, no test provided this information in the review by McCauley and Swisher (1984). Failure to conduct concurrent validity analysis for one other test (HAPP-3) may be somewhat understandable; obtained scores on the HAPP-3 are based only on analysis of phonological processes rather than accuracy of individual speech sounds, which would make the comparison of scores between the HAPP-3 and other tests very difficult to interpret. Although the overall finding for this criterion represents an improvement since 1984, the lack of information for one other test (GFTA-2) may be a particular concern because it appears to be one of the most widely used tests of this type (Skahan et al., 2007). That said, the popularity of the GFTA-2 is reflected in the fact that five other test developers (AAPS-3, CAAP-2, DEAP-A, LAT, SPAT-D II) compared their test to the GFTA-2 in their concurrent validity analyses. Concurrent validity data were also available in an independent study (Ogburn, 2008) that compared performance on the GFTA-2 with the AAPS-3. On a related note, however, Ogburn reported that the scores on the GFTA-2 were higher than on the AAPS-3. A similar report by the authors of the LAT (Bowers & Huisingh, 2010) in comparing that test to the GFTA-2 raises concerns about whether children might be underidentified with the GFTA-2. Further study is clearly indicated.

None of the tests presented findings on predictive validity. It is interesting that the authors of the CAAP-2 claimed to present predictive-validity data. However, a close inspection revealed that they had examined how well a group of children receiving speech services would have been classified as having a speech sound disorder using their test (reported as sensitivity and specificity). No actual prediction about future performance was being made. This was therefore judged to constitute a measure of diagnostic accuracy for which they were given credit herein. That said, the current finding of no true predictive-validity reports represents an identical outcome to that of McCauley and Swisher (1984). Although perhaps disconcerting, the author of one of the tests examined herein (Fudala, 2000) pointed out that expecting such data to be included may place an unfair burden on test developers. Indeed, from a practical perspective, waiting for such studies to be completed might result in a test whose stimuli and norms were outdated before the test had been published. From an ethical perspective it would also be problematic because conducting such studies in a meaningful way would require withholding intervention from at least some participants (i.e., the intervention would itself be a confound affecting the ability to make a prediction about future performance).

Relative to diagnostic accuracy, data were available for six of 10 tests. However, only two tests (CAAP-2, DEAP-A) provided sensitivity and specificity data. Thus, they were the only ones to offer detailed insight into how accurately the test both identified those with speech sound disorders and excluded those with typically developing speech. Four other tests did provide comparisons of test performance between typically developing children and children with speech sound disorders. However, the latter reports provide limited insight into how well these tests yield a valid diagnosis for any given child. Combined with four tests offering no

information on diagnostic accuracy, it raises serious concerns about our reliance on these instruments to make assessment and service eligibility decisions. Such concerns are of course consistent with the long-held view that clinicians should not rely exclusively on a single data point (i.e., a single test score) to determine whether intervention is needed.

The various test manuals addressed dialect and the range of variation between and within dialects to differing degrees. Four of the tests did not address the issue of dialect variation at all. Five (CAAP-2, BBTOP, DEAP-A, HAPP-3, and SPAT-D II) made a point of saying that clinicians need to be aware of dialect differences that could affect scoring and interpretation of the test results. Of these, two of the tests recommended that examiners be aware of dialect variation in their area and either take it into consideration during testing (CAAP-2) or create local norms (BBTOP). Two of the tests provided information on how to track dialect-based variation on the test protocol (DEAP-A, HAPP-3). One test manual (AAPS-3) stated that dialect adjustments were not needed because an independent study of an earlier edition of the test did not find that dialect adjustment (MAE vs. AAE) made a significant difference in test scores for a group of AAE speakers in the Detroit, Michigan, region (Washington & Craig, 1992). However, specific consideration was not given relative to the current edition of the test, and the study being cited was limited in scope to the variety of AAE spoken in Detroit, Michigan. Developers of one test (SPAT-D II) provided reference materials for clinicians on dialect difference to assist in test interpretation by presenting tables of phonological variation for several dialects or dialect groups, including AAE, Spanish-influenced English, and Asian-influenced English. Although test developers are beginning to consider dialect, this remains an area of concern. Due to the potential of misdiagnosing a speech difference as a disorder, this is a factor that clinicians must take into account. It may also influence their test selection.

Another dialect issue for future consideration is that none of the existing tests appeared to take into account potential differences in the course of development of speech milestones as well as in the development of phonotactic constraints, both of which were found to exist in a comparison of MAE and AAE (Pearson et al., 2009). For this reason, as well as the geographic variations that exist between speakers of the same or similar ethnic dialects, establishing local norms may be the best way to ensure appropriate test interpretation for speakers of nonmainstream dialects.

Only three of the tests reviewed included formal consideration of possible vowel errors. Test developers do not yet appear to have caught up with the field in terms of the need for this type of analysis. Six of the tests reviewed included the possibility of analysis of phonological patterns using the natural phonology framework. The number rises to seven if the GFTA-2 is included, wherein productions from that test can be analyzed using an independent procedure called the Khan–Lewis Phonological Analysis (Khan & Lewis, 2002). The inclusion of the natural phonology framework in many of these tests likely reflects the desire of clinicians to evaluate broader error patterns, which has long been thought to offer the potential of making intervention more efficient (Elbert & Gierut, 1986; Hodson & Paden, 1991; Williams, 2000).

For both test–retest and interexaminer reliability, eight of the test manuals included findings, but in both cases, fewer than half of the tests reported that all correlation coefficients were at least .90. For each criterion, several of the exceptions included reports of multiple values, with only some of the correlations being at least .90. Given that none of the tests examined by McCauley and Swisher (1984) met their criterion, the current findings would appear to represent improvement. That said, the failure of several of the tests to report values of at least .90 raises two concerns. First, it raises questions about the consistency of administration and scoring of these tests. This seems unlikely to have been the fault of the test developers, however, because instructions for administration and scoring were judged by the current authors to be sufficiently clear in all cases. In addition, all but one of the test developers specifically stated examiner qualifications. The second possibility is that the .90 criterion may itself be too high. All of the reported correlations across all the tests were .70 or higher, and the majority were .80 or higher. In this regard, lowering the criterion to .75 or higher would have resulted in seven of 10 tests meeting the criterion for both types of reliability. The mixed picture for reliability findings among the current tests may also arise from the fact that test developers did not all approach reliability measurement in the same way. Test manuals variously reported reliability coefficients for overall scores (e.g., AAPS-3), for both raw and standard scores (e.g., CAAP-2), for specific subtests (e.g., DEAP-A), for different age groups (e.g., LAT), or across speech production targets (e.g., GFTA-2). It is not at all clear whether the .90 cutoff is reasonable for all of these approaches to reach reliability.

Administration instructions were judged to be adequate for all 10 tests. Although the level of detail varied, all appeared to provide sufficient guidance so that the test could be reliably administered. Compared to the 4/5 test outcome for this variable reported by McCauley and Swisher (1984), this remains an area of strength for the tests in this area.

Nine of the 10 test manuals specified examiner qualifications, with the 10th (SHAPE) providing strong hints. This represents a significant improvement over the findings of McCauley and Swisher (1984), who reported that none of the tests they examined did so. However, the nature of the qualifications being suggested in the currently available tests warrants some consideration. Although some tests, such as the CAAP-2 and DEAP-A, limit their use to SLPs (or supervised trainees), most did not. Instead, most simply discuss training and preparation to varying degrees. For example, one manual (PAT-3) stated that examiners "should have some formal training in articulation assessment" (Lippke et al., 1997, p. 3). This contrasts with the somewhat more specific HAPP-3 manual, which states that

examiners "should have received instruction in phonetics, phonology, and diagnostic evaluation procedures and have experience administering speech sound assessment instruments" (Hodson, 2004, p. 5). The BBTOP indicated that the screening tool can be used by anyone able to accurately judge pronunciation accuracy at the single-word level. However, it specified that the test "was developed for use by speech-language clinicians" (Bankson & Bernthal, 1990, p. 2) and went on to say that examiners must also have skill in phonetic transcription and identification of phonological processes. The GFTA-2 manual tempered the qualifications question by discussing different levels of analysis that might be conducted (i.e., detecting the presence of an error, identifying the type of error, making judgments about intervention and prognosis). Different qualifications might be sufficient at different levels, though it is noted that "only those with appropriate training in speech pathology should make decisions about intervention and prognosis" (Goldman & Fristoe, 2000, p. 6). One test manual (LAT) even set limits on who can use it by specifically excluding paraprofessionals or support personnel. Two related issues arise here. The first is whether clinicians considering use of a particular test would consider themselves qualified to do so. That would appear to remain an individual decision. The second is the broader issue of whether as a field, specific training, experience, and/or qualifications should be mandated for using these tests. As noted previously, the present authors prefer to remain neutral in this regard. It may, however, be a question for our professional organizations to consider.

Relative to the last of the reliability-related criteria, eight of the tests reviewed included *SEM* data. This may well broaden the utility of our tests in this area. As noted by Hutchinson (1996), *SEM* data provide clinicians with a level of confidence around any individual obtained score. As discussed earlier, it also provides greater flexibility in test score interpretation as well as a reference point for judging test precision.

The normative samples were judged to have been adequately described for six of 10 tests reviewed. This represents a noticeable improvement over the findings of no tests meeting this criterion by McCauley and Swisher (1984). Three tests in the current review (BBTOP, PAT-3, SPAT-D II) failed the criterion because the test developers failed to include information on the SES of their sample. The need for such information remains important. Studies of the influence of SES on speech sound disorders have at best been mixed (see Bernthal et al., 2013, pp. 174–175), and as such, we cannot yet rule out the need to consider it in our assessment and intervention decisions.

Related to both the overall clarity of describing the normative sample and the next criterion, the developers of two tests failed to provide sufficient information about the sample to determine if it was limited to those who were typically developing or included those with speech sound disorders (i.e., if it was a truncated or a full-range sample). Thus, eight of 10 test manuals allowed for this classification. Six tests were classified as having full-range samples, one had a truncated sample, and one used a merger of both types of samples. Although the current report does not resolve the debate about which type of sample is most appropriate, there appears to be a trend among the tests for the use of full-range samples. Indeed, the one test (BBTOP) that appears to have used a truncated sample was the oldest test examined.

Relative to the size of the normative subgroups, three of the tests reviewed included at least 100 children in each subgroup from which scores are derived (three others met this criterion for the majority of their subgroups). This represents an improvement over the finding from McCauley and Swisher (1984) that none of the tests they reviewed did so. Despite the improvement, clinicians should continue to be cautious about using those tests that failed to meet this standard because they would potentially be making comparisons against relatively small normative groups. The necessary information is not always clear in the description of the normative sample. For example, the GFTA-2 manual indicates that there were more than 100 children at each age level (i.e., 3-year-olds, 4-year-olds, 5-year-olds, etc.) in both the male and the female groups. Although this was true, in the scoring tables each age level was split into smaller age increments. Thus, subgroup sizes for the GFTA-2 were not clearly specified. With an overall sample size of 2,350 children and 49 subgroups (see Table 1), subgroup sizes would average fewer than 50 children. The size of the subgroups becomes even more important with tests that have separate male and female norms.

Mean and standard deviation data were presented for all the subgroups for all 10 of the tests reviewed. This again represents an improvement over the finding of two of five tests reviewed by McCauley and Swisher (1984). However, the same caveat about overall subgroups applies here. If the subgroups themselves were too small, clinicians should be cautious about using the mean and standard deviation data derived from those subgroups.

Last, the question of whether it is necessary to provide separate scoring for boys and girls in the realm of speech sound disorders is not fully resolved by the current findings. Eight of the 10 test developers appeared to have examined this question. Six concluded that it was necessary for at least part of their age range or for some of the analyses, whereas two reached the opposite conclusion. Such a mixed conclusion would appear to mirror the existing literature, which continues to present a mixed picture about gender differences in speech sound acquisition in general. Continued examination of whether or not gender differences exist in particular normative samples would seem to be a prudent approach.

Overall, there has been a marked improvement in the psychometric qualities of single-word standardized tests of speech sound disorders, although there continue to be areas in need of improvement. As more information about test characteristics becomes available in test manuals, clinicians can make more informed decisions about test selection than in the past. Careful examination of the psychometric characteristics of such tests may help determine which ones would best serve the children on their caseloads. We hope

that this review can provide a starting point for making such decisions. It is important to keep in mind, however, that even with strong psychometric characteristics, single-word tests of articulation are limited in the information that they provide. They may not provide an adequate representation of a child's speech production skills in functional and conversational contexts (H. B. Klein & Liu-Shea, 2009; Morrison & Shriberg, 1992). Furthermore, as pointed out by Eisenberg and Hitchcock (2010), these tests also do not provide an adequate sample from which to determine a child's phonetic inventory or select appropriate intervention goals. As with most standardized tests, they best indicate a child's performance on particular items in comparison to other children of the same age. By themselves, such tests do not diagnose a disorder or provide guidance for the specifics of intervention.

Conclusions

The current review has been largely quantitative in its perspective. With a few exceptions, it has only looked at whether specific types of information were presented; it has not made judgments regarding the specific quality of that information. With that in mind, nine of 10 tests in this review met at least five of the criteria set out by McCauley and Swisher (1984). Many of the tests also met some additional criteria. Overall, it would appear that the state of the art in testing for speech sound production in children has improved. However, given that most of the tests failed to meet several of the criteria, there remains additional room for improvement.

It is tempting to try to prioritize certain criteria over others and identify which criteria are most important and/or have the most impact on the appropriateness of a given test in order to assist clinicians in test selection. However, given that tests are administered to different people for different purposes and in a variety of circumstances, it not possible to outline a set hierarchy of the most important aspects of standardized tests. In fact, the most recent version of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) specifically counsels against ranking of criteria, stating that "No type of evidence is inherently preferable to others; rather, the quality and relevance of the evidence to the intended test use determine the value of a particular type of evidence" (p. 23). Thus, clinicians are encouraged to carefully consider their testing needs and use the information presented herein to help them evaluate how well the tests that they select will serve their intended purposes.

Acknowledgments

An earlier version of this review was presented as a poster at the November 2013 Annual Convention of the American Speech-Language-Hearing Association in Chicago, Illinois. Thanks to Jennifer Montzka for her assistance with data collection.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Baker, E., Croot, K., McLeod, S., & Paul, R. (2001). Psycholinguistic models of speech development and their application to clinical practice. *Journal of Speech, Language, and Hearing Research, 44,* 685–702. doi:10.1044/1092-4388(2001/055)

Ball, M. J., & Gibbon, F. E. (2002). *Vowel disorders*. Boston, MA: Butterworth Heinemann.

Ball, M. J., & Kent, R. D. (1997). *The new phonologies: Developments in clinical linguistics*. San Diego, CA: Singular.

Bankson, N. W., & Bernthal, J. E. (1990). *Bankson-Bernthal Test of Phonology*. Austin, TX: Pro-Ed.

Bernthal, J. E., Bankson, N. W., & Flipsen, P., Jr. (2013). *Articulation and phonological disorders: Speech sound disorders in children* (7th ed.). Boston, MA: Pearson Education.

Bharadwaj, S. V., & Assmann, P. F. (2013). Vowel production in children with cochlear implants: Implications for evaluating disordered speech. *Volta Review, 113,* 149–169.

Bowers, L., & Huisingh, R. (2010). *LinguiSystems Articulation Test*. East Moline, IL: Linguisystems.

Cole, P. A., & Taylor, O. L. (1990). Performance of working class African-American children on three tests of articulation. *Language, Speech, and Hearing Services in Schools, 21,* 171–176. doi:10.1044/0161-1461.2103.171

Dawson, J., & Tattersall, P. (2001). *Structured Photographic Articulation Test II Featuring Dudsberry*. DeKalb, IL: Janelle Publications.

Dodd, B., Hua, Z., Crosbie, S., Holm, A., & Ozanne, A. (2006). *Diagnostic Evaluation of Articulation and Phonology*. San Antonio, TX: Pearson.

Eisenberg, S. L., & Hitchcock, E. R. (2010). Using standardized tests to inventory consonant and vowel production: A comparison of 11 tests of articulation and phonology. *Language, Speech, and Hearing Services in Schools, 41,* 488–503. doi:10.1044/0161-1461(2009/08-0125)

Elbert, M., & Gierut, J. A. (1986). *Handbook of clinical phonology: Approaches to assessment and treatment*. San Diego, CA: College-Hill Press.

Ertmer, D. J. (2010). Relationships between speech intelligibility and word articulation scores in children with hearing loss. *Journal of Speech, Language, and Hearing Research, 53,* 1075–1086. doi:10.1044/1092-4388(2010/09-0250)

Eveleigh, K. F., & Warr-Leeper, G. A. (1983). Improving efficiency in articulation screening. *Language, Speech, and Hearing Services in Schools, 14,* 223–232.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26,* 77–92.

Fudala, J. B. (2000). *Arizona Articulation Proficiency Scale–Third Edition*. Torrance, CA: Western Psychological Services.

Fudala, J. B., & Reynolds, W. M. (1986). *Arizona Articulation Proficiency Scale–Second Edition*. Los Angelas, CA: Western Psychological Services.

Goldman, R., & Fristoe, M. (2000). *Goldman Fristoe Test of Articulation–Second Edition*. San Antonio, TX: Pearson.

Harvill, L. (1991). An NCME instructional module on standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33–41.

Hodson, B. W. (2004). *Hodson Assessment of Phonological Patterns–Third Edition*. Austin, TX: Pro-Ed.

Hodson, B. W., & Paden, E. P. (1991). *Targeting intelligible speech: A phonological approach to remediation* (2nd ed.). Austin, TX: Pro-Ed.

Hutchinson, T. A. (1996). What to look for in the technical manual: Twenty questions for users. *Language, Speech, and Hearing Services in Schools, 27,* 109–121.

Kaufman, N. (1995). *Kaufman Speech Praxis Test for Children*. Detroit, MI: Wayne State University Press.

Khan, L., & Lewis, N. (2002). *Khan-Lewis Phonological Analysis–Second Edition*. San Antonio, TX: Pearson.

Klein, E. S. (1996). Phonological/traditional approaches to articulation therapy: A retrospective group comparison. *Language, Speech, and Hearing Services in Schools, 27,* 314–323.

Klein, H. B., & Liu-Shea, M. (2009). Between-word simplification patterns in the continuous speech of children with speech sound disorders. *Language, Speech, and Hearing Services in Schools, 40,* 17–30. doi:10.1044/0161-1461(2008/08-0008)

Lippke, B. A., Dickey, S. E., Selmar, J. W., & Soder, A. L. (1997). *Photo Articulation Test–Third Edition*. Austin, TX: Pro-Ed.

Madison, C. L., Kolbeck, C. P., & Walker, J. L. (1982). An evaluation of stimuli identification on three articulation tests. *Language, Speech, and Hearing Services in Schools, 13,* 110–115.

Mathias, B. N. (2010). *Psychometric review of speech tests for preschool children: 25 years later* (Unpublished undergraduate honors thesis). The Ohio State University, Columbus.

McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49,* 34–42.

McFadden, T. U. (1996). Creating language impairments in typically achieving children: The pitfalls of 'normal' normative. *Language, Speech, and Hearing Services in Schools, 27,* 3–9.

McLeod, S. (2013). Speech sound acquisition. In J. E. Bernthal, N. W. Bankson, & P. Flipsen Jr. (Eds.), *Articulation and phonological disorders: Speech sound disorders in children* (pp. 58–113). Boston, MA: Pearson.

Mikucki, B. A., & Larrivee, L. (2006, November). *Validity and reliability of twelve child language tests*. Poster presented at the Annual Convention of the American Speech-Language-Hearing Association, Miami, FL.

Morrison, J. A., & Shriberg, L. D. (1992). Articulation testing versus conversational speech sampling. *Journal of Speech and Hearing Research, 35,* 259–273.

Mullen, R., & Schooling, T. (2010). The National Outcomes Measurement System for pediatric speech-language pathology. *Language, Speech, and Hearing Services in Schools, 41,* 44–60. doi:10.1044/0161-1461(2009/08-0051)

Ogburn, A. C. (2008, November). *A comparison of the GFTA-2 and the Arizona-3: A clinical focus*. Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Chicago, IL.

Pearson, B. Z., Velleman, S. L., Bryant, T. J., & Charko, T. (2009). Phonological milestones for African American English-speaking children learning mainstream American English as a second dialect. *Language, Speech, and Hearing Services in Schools, 40,* 229–244. doi:10.1044/0161-1461(2008/08-0064)

Peña, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology, 15,* 247–254.

Perona, K., Plante, E., & Vance, R. (2005). Diagnostic accuracy of the Structured Photographic Expressive Language Test: Third edition (SPELT-3). *Language, Speech, and Hearing Services in Schools, 36,* 103–115.

Preston, J. L., Brick, N., & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology, 22,* 627–643. doi:10.1044/1058-0360(2013/12-0139)

Secord, W., & Donahue, J. S. (2014). *Clinical Assessment of Articulation and Phonology–2*. Greenville, SC: Super Duper Publications.

Shanks, S. J., Sharpe, M. R., & Jackson, B. R. (1970). Spontaneous responses of first grade children to diagnostic picture articulation tests. *Journal of Communication Disorders, 3,* 106–117.

Shriberg, L. D., & Austin, D. (1998). Comorbidity of speech-language disorder: Implications for a phenotype marker for speech delay. In R. Paul (Ed.), *Exploring the speech-language connection* (pp. 73–117). Baltimore, MD: Brookes.

Shriberg, L. D., & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders, 47,* 256–270.

Skahan, S. M., Watson, M., & Lof, G. L. (2007). Speech-language pathologists' assessment practices for children with suspected speech sound disorders: Results of a national survey. *American Journal of Speech-Language Pathology, 16,* 246–259.

Smit, A. B., & Hand, L. (1997). *Smit-Hand Articulation and Phonology Evaluation*. Los Angeles, CA: Western Psychological Services.

Stampe, D. (1979). *A dissertation on natural phonology*. New York, NY: Garland Publishing.

Torrington Eaton, C., & Bernstein Ratner, N. (2013). Rate and phonological variation in preschool children: Effects of modeling and directed influence. *Journal of Speech, Language, and Hearing Research, 56,* 1751–1763. doi:10.1044/1092-4388(2013/12-0171)

Washington, J. A., & Craig, H. K. (1992). Articulation test performances of low-income, African-American preschoolers with communication impairments. *Language, Speech, and Hearing Services in Schools, 23,* 203–207.

Williams, A. L. (2000). Multiple oppositions: Theoretical foundations for an alternative contrastive intervention approach. *American Journal of Speech-Language Pathology, 9,* 282–288.

Williams, A. L., McLeod, S., & McCauley, R. J. (2010). *Interventions for speech sound disorders in children*. Baltimore, MD: Brookes.