A Review of Standardized Tests of Nonverbal Oral and Speech Motor Performance in Children

Rebecca J. McCauley
University of Vermont, Burlington

Edythe A. Strand
The Mayo Clinic, Rochester, MN

Purpose: To review the content and psychometric characteristics of 6 published tests currently available to aid in the study, diagnosis, and treatment of motor speech disorders in children.
Method: We compared the content of the 6 tests and critically evaluated the degree to which important psychometric characteristics support the tests' use for their defined purposes.
Results: The tests varied considerably in content and methods of test interpretation. Few of the tests documented efforts to support reliability and validity for their intended purposes, often when relevant information was probably available during the test's development.
Conclusions: Problems with the reviewed tests appear related to overly broad plans for test development and inadequate attention to relevant psychometric principles during the development process. Recommendations are offered for future test revisions and development efforts that can benefit from recent research in test development and in pediatric motor speech disorders.

Key Words: children, motor speech disorders, assessment, childhood apraxia of speech, dysarthria

Clinicians and researchers turn to a wide array of clinical measures to assist in decision making about diagnosis, treatment planning, and assessment of progress. These measures include informal probes, published checklists, and standardized, elicited samples of behavior. Minimally, measures can be considered standardized tests when they specify standard procedures for administration and interpretation (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999). To be considered well developed as well as standardized, however, tests must also be supported by evidence that they function as intended when used as recommended. Specifically, they must demonstrate evidence of reliability and validity for their intended purposes and population (AERA et al., 1999; Dollaghan, 2004; Haladyna, 2006; Linn, 2006; Messick, 1989, 1995). To date, however, only sporadic efforts in the form of individual test reviews have been made to examine the extent to which well-developed standardized tests are available for assessing speech and nonspeech oral motor function in children (e.g., Guyette, 2001; McCauley, 2003; Towne, 2001). A critical review of the content and psychometric characteristics of tests for children with motor speech disorders is necessary and overdue.

Reviewing a test's psychometric characteristics entails critically evaluating evidence that the test functions as intended (i.e., is valid) for the purposes and populations for which it was originally developed and is currently used. As part of such a review, evidence of reliability is evaluated as a necessary prerequisite for validity: A less reliable test is necessarily a less valid one. Other elements of the test's preliminary or ongoing development are also reviewed because they, too, indirectly support the argument for validity. For example, the quality of norms or behavioral standards linking test results to clinical decisions also affects the degree to which the test can be validly used for a specific purpose (Buckendahl & Plake, 2006; Cizek, 2006). Therefore, their description and justification are also routinely reviewed as part of an overall psychometric examination.

Three serious challenges seem likely to impede test authors' pursuit of high psychometric standards (e.g., AERA et al., 1999) for tests of motor speech production in children. These challenges are (a) the field's limited understanding of the nature of motor speech disorders in children, (b) the

1058-0360/08/1701-0081

changing manifestations of these disorders at differing stages of children's development, and (c) the difficulties associated with devising performance tests for young children. These three challenges present compelling reasons to believe that tests in this area may not be well developed and that taking stock of their current status can serve as an important step toward promoting their judicious use and improving the quality of ongoing test development in this area.

The field's limited understanding of motor speech disorders in children is demonstrated most powerfully by a lack of agreement on core characteristics that can help guide test construction and validation and lead to the development of a test that can serve as a gold standard. This lack of agreement is particularly evident for childhood apraxia of speech (CAS), for which many different core characteristics have been proposed (Caruso & Strand, 1999; Davis, Jakielski, & Marquardt, 1998; Forrest, 2003; McCabe, Rosenthal, & McLeod, 1998; Shriberg, Aram, & Kwiatkowski, 1997a). The recent publication of a position statement concerning CAS (American Speech-Language-Hearing Association, 2007) may help reduce this controversy somewhat (especially with regard to the existence of the disorder) but is unlikely to entirely eliminate controversies surrounding it. Although dysarthria is a more universally recognized category of pediatric motor speech disorders, there is still no gold standard for distinctions among dysarthria types during speech development. In addition, some types of dysarthria may be difficult to distinguish from apraxia, especially in children (Yorkston, Beukelman, Strand, & Bell, 1999).

The absence of a gold standard is a common problem facing researchers and clinicians dealing with a variety of behavioral and medical disorders (Aronowitz, 2001; Dollaghan, 2004; Feinstein, 2001; Streiner & Norman, 2003). It places special demands on test authors in their validation efforts because they cannot obtain evidence that the test takers' performances on the new test parallel those of a gold standard. Denied this relatively straightforward evidence of validity, authors must use more elaborate methods of validation. Absence of a gold standard also precludes the detailed description of a new test's diagnostic accuracy using metrics from clinical epidemiology that are increasingly used in speech-language pathology (e.g., sensitivity and specificity; Fletcher & Fletcher, 2005; Spaulding, Plante, & Farinella, 2006). These require a gold standard, or an arguable substitute for it, for their calculation (Dollaghan, 2004).

A second challenge facing test authors is the changing nature of children's motor speech disorders over time (Lewis, Freebairn, Hansen, Iyengar, & Taylor, 2004; Shriberg, Campbell, et al., 2003; Strand, 2002; Yorkston et al., 1999). Because of such change, tests that might be quite appropriate for use in diagnosis of the disorder at one age or level of severity may be quite inappropriate for that purpose at other ages and levels of severity. For example, whereas prosodic abnormalities may be observable in children with CAS who have attained a certain level of speech production skill, such abnormalities may not be apparent in children who are more severely affected and produce limited speech. Similarly, deficits in production of multisyllabic words may be particularly significant in some older children suspected of having CAS (Shriberg, Aram, & Kwiatkowski, 1997b), but such abnormalities may be very difficult to identify in younger and/or severely affected children who may produce only monosyllabic words or almost no speech at all.

A third challenge to test authors interested in children's motor speech disorders is the special difficulty posed by the development of performance tests, such as those involved in the evaluation of nonspeech oral motor and motor speech functions, in which behavior samples are elicited and evaluated (e.g., Robbins & Klee, 1987; Thoonen, Maassen, Wit, Gabreels, & Schreuder, 1996). Performance tests, in which open-ended responses and skilled evaluation of the response are the rule, have received less attention from psychometricians than conventional educational tests that use relatively constrained responses (e.g., multiple-choice formats) and scoring systems (e.g., right/wrong; AERA et al., 1999; Bennett, 1993; Welch, 2006). Such tests are known to place complex cognitive demands on test takers and evaluators alike (Cizek, 2006; Rvachew, Hodge, & Ohberg, 2005). Understandably, their development is especially difficult when maximal performance is sought (e.g., diadochokinetic rates) and when the test takers are young children whose attention, cooperation, and even understanding of task requirements are often uncertain (Davis & Velleman, 2000; Kent, Kent, & Rosenbek, 1987).

Based on the unexamined quality of current tests and the challenges posed in their development, there exists a pressing need to examine the content and psychometric characteristics of tests designed to assess speech and nonspeech oral motor function in young children. The purpose of this article is to provide such an examination. In it, we discuss content descriptively by comparing the intended populations and purposes of six tests developed to measure children's nonverbal oral motor and/or motor speech skill. We then evaluate their psychometric adequacy using operational definitions based on traditional expectations for norm-referenced measures, criterion-referenced measures, or both, depending upon the author's intended purpose (AERA et al., 1999; Buckendahl & Plake, 2006; McCauley, 2001). In particular, reliability and validity are examined. We conclude with a summary of the current state of these tests as well as recommendations for the future use and development of tests in this area.

Method

Test Search Strategy

We located candidate tests published between January 1990 and July 2006 through searches of publisher catalogs received in the first author's academic department and three additional sources: the Health and Psychosocial Instruments (HaPI) database, the Buros Institute's text yearbooks (e.g., Plake & Impara, 2001), and the Buros Institute's Test Reviews Online (Buros Institute, 2006). In publisher catalogs, we examined entries under all of the following categories: *assessment or evaluation, motor speech, articulation,* and *phonology.*[1] Key

[1]Catalogs were from the following companies: AGS/Pearson Assessment, Brookes, Harcourt Assessment, Janelle, LinguiSystems, Pro-Ed, Riverside, Thinking Publications, Wayne State University Press, and Western Psychological Services.

words used in the electronic search of the HaPI database were *speech or verbal or oral, motor or praxis or apraxia or execution,* and all words containing the root *child*. To search the two Buros Institute sources, we looked within the following preexisting categories: *speech and hearing, neuropsychological,* and *sensory-motor.*

Inclusionary/Exclusionary Criteria

We selected tests for review if they (a) were standardized (i.e., included standard stimuli and instructions for administration and interpretation); (b) included young children (i.e., children at or below elementary school age) among the age groups for which the measure was intended; (c) addressed nonverbal oral motor or motor speech function, or both; and (d) were available in July 2006 through a commercial source. From these, we excluded tests if they focused solely on oral mechanism structure (e.g., the Oral Mechanism Examination for Children and Young Adults; Riski & Witzel, 2001) or sound system analyses (e.g., the Hodson Assessment of Phonological Patterns—Third Edition; Hodson, 2003). These criteria were used in order to focus on tests most likely to be used to answer questions related to the management of the special speech production challenges of children with motor speech disorders.

The Review Process

The first author reviewed selected tests for the following information: (a) information about the population for whom the test is intended, (b) purposes for which the test is intended to be used, (c) item content, (d) norms and/or behavioral standards used to guide score interpretation, and evidence of (e) reliability and (f) validity. Information concerning each test's target populations and purposes was sought in introductory sections of the manual and with sections on test administration and interpretation.

Categorization of test content. We categorized the content of each test item as assessing (a) nonverbal oral motor function, (b) motor speech function, or (c) oral structure. We categorized items as assessing *nonverbal oral motor function* when movements of the jaw, lips, tongue, or palate were observed in nonspeech contexts; *speech motor function* when movements of articulators were examined during speaking; and *oral structure* when the structures were examined at rest. (Although we had not planned to address oral structure content in this study, we did so in order to describe the content of the reviewed tests comprehensively.) We calculated percentages of items within each category as the number of items in that content area divided by the total number of items in the test, multiplied by 100.

Within the larger category of nonverbal oral motor function, we further subcategorized items according to nonfeeding or feeding functions. Items were subcategorized as assessing *nonfeeding oral motor functions* if they involved movements outside of the context of both speech production and food (e.g., "Show me how you blow"; Hayden & Square, 1999, p. 44). They were subcategorized as assessing *feeding* if they directly or indirectly assessed oral movement patterns in relation to food (e.g., "Child displays the ability to easily bite through various food thicknesses which are age appropriate"; Jelm, 2001, p. 11). We calculated percentages of each subcategory as the total number of items in the subcategory divided by the total number of nonverbal oral motor items on the test, multiplied by 100.

Evaluation of psychometric characteristics. We evaluated psychometric characteristics of each test by examining the methods the test authors specified for interpreting test takers' performances as well as for demonstrating the test's reliability and validity. Table 1 summarizes the operational definitions we used to judge adequacy of psychometric characteristics, which are described in greater detail in this section.

Rather than making binary judgments of adequacy, we applied operational definitions using a three-way distinction to evaluate the methods used by test authors to guide users in test interpretation and to support claims of reliability and validity. Specifically, we made a distinction between a test's (a) providing no relevant information about the characteristic being examined, (b) providing some information but failing to meet the operational definition for adequacy, and (c) meeting the operational definition for adequacy. Information about earlier versions of the test was examined, but it was not considered relevant to determining adequacy for the current version of the test.

First, we examined the method specified for the interpretation of test performance. For each test, we examined the quality of norms, behavioral standards used in test interpretation, or both, based on each test's intended purposes (Linn, 2006; McCauley, 1996, 2001). Specifically, we examined *norms* for tests that indicated use for diagnosis or screening as purposes and *behavioral standards* for tests that indicated treatment planning and/or examining behavioral change as purposes (McCauley, 1996, 2001; Merrell & Plante, 1997).

To judge the adequacy of test norms, we examined test authors' methods for normative group identification and description. In order to meet the operational definition for adequate test norms, authors needed to indicate the methods by which they determined group membership (e.g., testing, parental and teacher report, previous diagnosis) and to describe characteristics of group members that would help test users judge the applicability of the norms to their client. For groups with typical speech, these descriptive characteristics were age, gender, and the presence (or absence) of nonspeech difficulties. For groups with disordered speech, a judgment of adequacy required an additional fourth characteristic—the severity of the speech disorder.

To judge the adequacy of behavioral standards used for decision making, we evaluated test authors' methods for identifying and justifying any behavioral standard used in test interpretation. For example, when a manual instructed test users to plan treatment based on profile analysis, the test developer needed to specify how performance on a specific subtest or item(s) would be determined to be worthy of treatment focus (e.g., by specifying the cut score at which a subtest would be considered a reasonable treatment focus). Alternatively, when a manual instructed test users to use the test to examine changes in a child's skills over time, the test developer needed to specify how significant change (rather than change due to test error) would be identified and whether the test user was to look for such change for the test as a

TABLE 1. Psychometric characteristics examined in this review.

Characteristic	Operational definition for judgment of adequacy
Method of interpreting test performance	
Comparison to norms	For all normative groups, all of the following were needed to meet the operational definition: (a) Specification of methods used to assign children to groups (b) Description of subgroups in terms of age, gender, and co-occurring problems For normative groups with speech disorder, the following additional element was required: (c) Description of subgroups in terms of severity of speech disorder
Comparison to a behavioral standard	For all standards used to address either treatment planning or assessing change over time, both of the following were needed to meet the operational definition: (a) Specification of all behavioral standards (b) Justification of the standard(s)
Reliability	
Test–retest	All of the following were needed to meet the operational definition: (a) Report of a reliability study with a statistically significant correlation coefficient of .90 or higher (b) Clear description of the study participants (c) Specification of the time period between test administrations
Interexaminer	All of the following were needed to meet the operational definition: (a) Report of a reliability study with a statistically significant correlation coefficient of .90 or higher (b) Clear description of the study participants (c) Clear description of examiner qualifications
Validity	
Content	Any of the following were needed to meet the operational definition: (a) Description and justification of methods used to choose content, including a discussion of content relevance and coverage (b) Report of expert evaluation of test content (c) Use of an item analyses, including a description of study participants and statistical methods
Criterion-related	One or more study in which test scores correlated, as predicted, with a second well-motivated measure of the test's underlying construct. The study participants and statistical methods were described.
Construct	Any of the following were needed to meet the operational definition: (a) Evidence from a factor analytic study confirming expectations of the test's internal structure (b) Evidence that test performance improves with age (c) Evidence that groups that were predicted to differ in test performance actually do so. In addition, evidence needed to be obtained within a study in which statistical methods and participants were described.

whole, for subtests, or individual items. We examined sections of the test manual dealing with test interpretation for this information.

We then turned to examination of reliability and validity. With regard to reliability, we examined (a) test–retest and (b) interexaminer reliability. For both types of reliability, the magnitude and statistical significance of the reliability coefficient and the quality of the study in which it was obtained were of interest. Any reliability correlation coefficient needed to be statically significant and at least .90, a magnitude that is frequently cited as a minimum standard for diagnostic purposes (Salvia & Ysseldyke, 2007). Whereas the size of the correlation is deemed important for describing the degree of relationship, statistical significance was also evaluated in order to rule out findings that may have been due to chance. In addition, we examined the study in which the correlation data were obtained. In order to meet the operational definition for adequacy, the study needed to include adequate description of participants and testing intervals for test–retest reliability, or description of participants and examiner qualifications for interexaminer reliability. We sought information related to this operational definition in the reliability section of each test manual.

With regard to validity, we examined three categories of evidence: *content validation* (evidence that the test content is relevant and adequately covers the construct being assessed), *criterion-related validation* (evidence that the test performs in a manner similar to another measure that is thought to be a valid indicator of the construct), and evidence of *construct validation* (evidence that the test functioned as predicted, given the assumption that it successfully measures the underlying construct). We looked for all forms of validity evidence under sections of test manuals related to validity, but also examined the description of test development for evidence of content validation.

The operational definitions for the three sorts of evidence used for test validation were intended to provide considerable flexibility for how test authors approached that process. For content validation, the test manual needed to provide *any* of the following in order to be rated as adequate: (a) description and justification of methods used to choose content so that it was relevant and represented an adequate sample of the construct being assessed, (b) expert opinions to verify that test content appeared relevant and comprehensive, or (c) an item analysis to study items during the test's development. For adequacy of criterion-related evidence of validity, the test manual needed to describe one or more studies in which test takers' performances on the test correlated with an alternative, well-motivated measure of the construct. For construct validation, the test manual needed to provide evidence of any of the following: that the relationships of performances on different items within the test met expectations concerning

TABLE 2. Age ranges and purposes of the reviewed tests.

Test	Age range (years;months)	Purposes			
		Screening	Diagnosis	Treatment planning	Assessing change over time
AP	3;0–13;11		✓	✓	
KSPT	2;0–6;0		✓	✓	
OSMSE–3	5;0–77	✓			
STDAS–2	4;0–7;11	✓			✓
VDP	Not specified		✓	✓	
VMPAC	3;0–12;0		✓	✓	✓

Note. Checkmark indicates that the test was described as appropriate for that purpose. AP = Apraxia Profile; KSPT = Kaufman Speech Praxis Test for Children; OSMSE–3 = Oral Speech Mechanism Screening Examination, Third Edition; STDAS–2 = Screening Test for Developmental Apraxia of Speech—Second Edition; VDP = Verbal Dyspraxia Profile; VMPAC = Verbal Motor Production Assessment for Children.

the test's internal structure (a factor analytic study), that test performances improved with age (a developmental study), or that groups expected to differ on the test actually did so (a contrasting groups study; McCauley, 2001). In addition, the research studies used to provide any validity evidence needed to be described both in terms of participant characteristics and statistical methods.

Results

Only 6 of the 22 tests identified for possible review met inclusionary and exclusionary criteria. These were the Apraxia Profile (AP) Preschool and School-Age Versions (Hickman, 1997); the Kaufman Speech Praxis Test for Children (KSPT; Kaufman, 1995); the Oral Speech Mechanism Screening Examination, Third Edition (OSMSE–3; St. Louis & Ruscello, 2000); Screening Test for Developmental Apraxia of Speech—Second Edition (STDAS–2; Blakeley, 2001); the Verbal Dyspraxia Profile (VDP; Jelm, 2001); and the Verbal Motor Production Assessment for Children (VMPAC; Hayden & Square, 1999). The remaining 16 tests were omitted from the review because they did not include focus on nonverbal oral motor and/or motor speech performance.[2]

Table 2 summarizes two characteristics for the six reviewed tests that significantly affect any test's relevance for specific children and assessment questions: (a) age range covered by the test and (b) the test's intended purposes. The manual for the VDP did not specify an appropriate age range, but manuals for the other tests advocated use with children of widely varying ages as well as with adults. Only the age range specified for the KSPT included children below the age of 3 years.

Stated purposes of the six tests included diagnosis, screening, treatment planning, and measuring change over time. Authors also stated use as a research measure and use as a training tool as purposes. Because the last two of these "uses" specify a context rather than a specific use, they were not included in Table 2. Authors of five of the six tests (all except the OSMSE–3) described them as appropriate for multiple purposes. All six of the test authors endorsed their test's use in either screening or diagnosis—purposes that usually involve norm-referenced interpretation or an empirically derived cutoff based on a group comparison (McCauley, 2001; Merrell & Plante, 1997; Pena, Spaulding, & Plante, 2006). Authors of five tests indicated that they were appropriate for purposes that are usually associated with a criterion-referenced mode of interpretation, such as treatment planning or assessing behavioral change over time (McCauley, 1996, 2001). In such cases, the specification of a behavioral standard serves a similar interpretative function as test norms for a norm-referenced interpretation.

Test Content

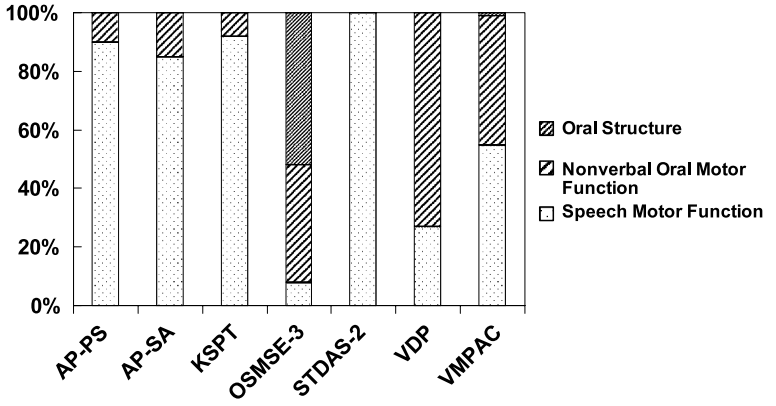
Despite the reviewed tests' overlapping purposes, their content varied considerably (see Figure 1). Oral structure was assessed in only two tests: the VMPAC (1% of test items) and the OSMSE–3 (50% of test items). Nonverbal oral motor function was assessed in five of the six tests, but composed anywhere from 8% (KSPT) to 73% (VDP) of each of those tests' items. Motor speech function was assessed in all tests and composed from 8% (OSMSE–3) to 100% (STDAS–2) of the test's content. Thus, although five tests included both nonverbal oral motor and speech motor content (excluding only the STDAS–2), the relative proportion of items devoted to these content areas varied considerably across tests.

Within the five tests that sampled nonverbal oral motor function, both feeding and nonfeeding items were included in some tests. Items related to nonfeeding were more common than those related to feeding. Two tests (KSPT and OSMSE–3) included no items related to feeding, whereas two others had some items related to feeding: the AP (Preschool form, 33%, and School-Age form, 15%) and VMPAC (7%). In contrast, the VDP included more feeding than nonfeeding items (58% vs. 42%, respectively).

Although motor speech items were included in all six tests, the items varied considerably in complexity, task requirements, and type of judgment made by the tester. Items were

[2]Aase et al., 2000; Bankson & Bernthal, 1990; Dawson & Tattersall, 2001; Fisher & Logemann, 1971; Fudala & Reynolds, 2000; Goldman & Fristoe, 2000; Hodson, 2003; Khan & Lewis, 2002; Lanphere, 1998; Masterson & Bernhardt, 2002; Pendergast, Dickey, Selmar, & Soder, 1997; Riski & Witzel, 2001; Secord, 1981; Secord & Donohue, 1998; Weiss, 1980; Wilcox & Morris, 1999.

FIGURE 1. Percentage of items in three content areas for the reviewed tests, including both Preschool (AP-PS) and School-Age (AP-SA) versions of the Apraxia Profile (AP). KSPT = Kaufman Speech Praxis Test for Children; OSMSE–3 = Oral Speech Mechanism Screening Examination, Third Edition; STDAS–2 = Screening Test for Developmental Apraxia of Speech—Second Edition; VDP = Verbal Dyspraxia Profile; VMPAC = Verbal Motor Production Assessment for Children.



usually organized according to linguistic or motor speech complexity (e.g., single phoneme production, syllable production, word production, sentence production). Methods of elicitation varied considerably (e.g., imitated vs. spontaneous, number of repetitions). Tests also used varying methods of manipulating complexity. Accuracy was measured for numerous aspects of the child's performance (e.g., prosody, phonetic accuracy, and syllable sequencing), but varied widely across tests.

Psychometric Characteristics

Methods of interpreting test performance: norms. We considered the quality of norms for all six tests because their test manuals indicated diagnostic and/or screening purposes. Table 3 indicates that only half of the tests provided information on norms.

The VMPAC was the only test we reviewed that met our operational definition for adequate norms. Although the

TABLE 3. Adequacy of methods of interpreting test performance, reliability, and validity for the reviewed tests.

Method of interpreting test performance	AP	KSPT	OSMSE–3	STDAS–2	VDP	VMPAC
Comparison to norms	—	*	—	*	—	✓
Comparison to a behavioral standard						
Treatment planning	*	*	NA		*	*
Assessing change	—	*	NA	—	—	—
Reliability						
Test–retest reliability	—	*[a]	—	—	—	*[b]
Interexaminer reliability	—	—	—	—	—	*[c]
Validity						
Content	—	—	—	*[d]	—	✓[e]
Criterion-related	—	*[f]	—	—	—	—
Construct	—	—	—	*[g]	—	*[h]

Note. Performance was rated as follows: No relevant information was provided (—); failed to meet operational definition for adequacy, but some relevant information was provided (*); met the operational definition for the current version of the test (✓); not applicable (NA). Footnotes for the reliability evidence indicate subtests for which evidence was offered. Footnotes for the validity evidence indicate types of evidence offered.

[a]Oral Movement and the Simple Phonemic and Syllabic Level subtests.
[b]Focal Oromotor Control.
[c]Global Motor control, Focal Oromotor Control; Sequencing, Connected Speech and Language Control; Speech Characteristics.
[d]Content justification and Item analysis.
[e]Content justification.
[f]Correlation study.
[g]Developmental study and Contrasting groups study.
[h]Developmental study and Contrasting groups study.

VMPAC manual reported on four clinical groups, only data for the single group of children with typical speech were used as norms. For the normative group, children's eligibility was based on age, English language use as determined by parent and examiner report, and ability to take the VMPAC with no test protocol modifications. Children were not excluded based on other disabilities. The normative group as a whole was described in terms of percentages of age, gender, and "an overall level of attention deficit disorder, developmental delay, emotional disturbance, learning disability, or other health impairment," as well as in terms of "gifted and talented status" (Hayden & Square, 1999, p. 15).

Two tests, the KSPT and the STDAS–2, used two normative groups—one with typical speech and one with disordered speech. The KSPT test manual included information about how children were selected for each of those groups but did not appropriately describe each of the groups and so did not satisfy the operational definition used for adequate description. In particular, no information was offered concerning the presence or absence of nonspeech problems for the 447 children with typical speech. Neither was the severity of the speech impairment described for the 263 children composing the clinical group. The STDAS–2 also did not meet the operational definition for adequate norms; specifically, it did not provide information about nonspeech problems in either the group of 51 children with normal speech or the group of 49 children with CAS. In addition, children selected for the group with CAS were identified by referral because they were "suspected of or identified as having developmental apraxia of speech" (Blakeley, 2001, p. 13); however, the test manual did not mention any additional steps confirming diagnosis, and no description of severity was provided.

Methods of interpreting test performance: behavioral standards. Manuals for five of the six reviewed tests advocated their use for one or more purposes usually associated with criterion-referenced interpretation of test performance (e.g., planning treatment). Therefore, an examination of the quality of behavioral standards is as relevant for them as examination of the quality of norms is for tests used for screening or diagnosis. Whereas the VDP manual specified its use for treatment planning only, manuals for four tests (AP, KSPT, STDAS–2, and VMPAC) indicated they could be used for both treatment planning and assessing change over time.

None of the five tests indicating treatment planning as an intended use specified a behavioral standard to serve as a basis for deciding whether specific areas of content should be addressed in treatment. The KSPT and VMPAC manuals indicated that examining performance across *subtests* would support treatment planning, but they went no further in describing the method to be used. The KSPT manual provided an example in which performance at 1 *SD* below the mean of the typical group's performance on a subtest was the apparent basis for recommending treatment on related content. However, the manual contained no explicit statement of that cut-score as a standard and provided no rationale for the implied cutoff. The VMPAC manual provided five case examples of children with problems of varying severity, but no explicit standards and consequently no justification for them. Even less guidance was provided by manuals for the AP, STDAS–2, and VDP. These tests indicated that errors on test *items* should be used to guide treatment planning but provided no guidance about how that purpose should be accomplished and no rationale for this method.

Examination of change over time was listed as an appropriate use by manuals for the AP, KSPT, STDAS–2, and VMPAC. Nonetheless, the AP, STDAS–2, and VMPAC manuals did not indicate how test performance could be used to help achieve this purpose. The KSPT manual contained the statement that "Progress can be quantified in several ways" (Kaufman, 1995, p. 3), then noted the availability of raw scores and derived scores (percentiles, age equivalents) associated with norms for two groups with typical and disordered speech, respectively. However, it provided no additional information concerning what would constitute a meaningful demonstration of change (i.e., it offered no behavioral standard for distinguishing measurement error from significant change). In summary, none of the five test manuals explicitly stated and justified a behavioral standard for assessing change.

Reliability. Only the KSPT and the VMPAC manuals provided information about test–retest reliability, which was reported for individual sections or subtests. Based on a reliability study of children from the disordered speech group, the KSPT obtained test–retest reliability coefficients for two subtests (the Oral Movement and the Simple Phonemic and Syllabic Level subtests) that met or exceeded our required level of .90 for adequacy of reliability but were not described in terms of statistical significance. Further, children in the disordered speech group were not described in terms of age or severity, and the test–retest interval was unspecified. The reliability study reported for the VMPAC used a well-described group of 115 children over a 7- to 14-day test–retest period. A reliability coefficient greater than .90 was obtained for the Focal Oral Motor Control section of the test, but the manual did not indicate whether this coefficient was statistically significant. Therefore, the Focal Oromotor Control section of the VMPAC did not achieve the operational definition simply because the question of statistical significance was not addressed.

For current versions of the tests, interexaminer reliability data were only reported for the VMPAC. All five VMPAC subtests yielded reliability coefficients that met or exceeded .90, but no information was provided about the statistical significance of these correlations. These correlations were obtained in a study in which 119 children from the standardization sample were simultaneously tested by one examiner and scored by another. Although the children were described in terms of age and racial/ethnic background, the number and training of examiners who participated in this study were not stated. Therefore, the VMPAC did not provide all of the information required to meet the operational definition for adequate interexaminer reliability.

Validity. For evidence of validity, the operational definitions allowed for several different kinds of evidence (see Table 1). Only two tests—STDAS–2 and VMPAC—provided information regarding content validation for current versions of the test. Both did so by attempting to offer credible justifications of item content and item analysis methods. The

VMPAC offered justifications of item content, thus allowing it to meet the operational definition for content validation. The STDAS–2 did not because its justification of items was incomplete. Neither the VMPAC nor STDAS–2 provided convincing evidence related to item analysis.

Only the KSPT pursued criterion-related validation using a current version of the test. In that validity study, the test developer correlated three subtests (Oral Movement, Simple Level, and Complex Level) with ratings of spontaneous speech for the two groups of children from the normative sample (typical speech and speech disordered). However, none of the resulting correlations reached the .90 level specified in the operational definition. Therefore, this method of validation was not successfully pursued by any of the six tests we evaluated.

Construct validation methods were reported for current versions of two tests (STDAS–2 and VMPAC) out of the six we examined. Both attempted to provide such evidence using contrasting group and developmental studies. Neither test manual described studies that satisfied the operational definition for this type of evidence. The STDAS–2 contrasting groups studies provided insufficient information about participants and no statistical analysis. The STDAS–2 developmental study, which was associated with a moderate correlation between age and performance on one subtest (Articulation), also did not adequately describe participants. The VMPAC provided tabular and graphic presentations of mean data to argue for evidence of developmental trends as well as for evidence of differences across contrasting groups. Although the groups were quite well defined in terms of demographic and other characteristics in each study, statistical examination of the data was missing.

Summary of Results

Although test manuals offered some evidence related to norms and behavioral standards for use in test interpretation, few documented efforts to support reliability and validity. Only the VMPAC provided norms that were adequately described. None of the tests provided clearcut behavioral standards on which to base decisions regarding treatment planning or change in performance over time. The VMPAC came closest of any of the tests to meeting operational definitions for the adequacy of its reliability information but did not meet them due to a lack of statistical detail. The VMPAC was also the only test to meet any of the three operational definitions for validation (i.e., content validation). Across tests, the absence of information about participants and a lack of attention to statistical support for evidence were largely responsible for unmet operational definitions.

Discussion

The purpose of this study was to review the content and psychometric characteristics of the six published standardized tests currently available to aid in the study, diagnosis, and treatment of motor speech disorders in children. Although tests in this area were found to be inadequately developed from a psychometric perspective, some probable sources of their deficiencies were identified. How to deal with these limitations in the present and avoid them in the future are topics meriting serious discussion.

Looking at the test content and methods of score interpretation undertaken by test authors, we saw considerable and possibly perilous complexity. All but one of the tests addressed two or more major content areas (motor speech function, nonverbal oral motor function, oral structure), and all addressed at least two purposes (screening, diagnosis, planning treatment, assessing behavior change). Because these content areas and purposes are recognized as important to a comprehensive assessment of motor speech disorders (Strand & McCauley, 1999), it is understandable that test authors would attempt to develop a single test to address them. Doing so would seemingly promise potential efficiencies for the test user and developer alike. However, addressing multiple content areas and purposes poses additional demands on the test development process. For example, items designed to be used for diagnosis versus treatment planning are often written and selected on differing grounds, resulting in diverging strategies of test development. These complications are possibly not the best use of limited resources of time and capital. Consequently, those undertaking future development or revision of tests in this area should either limit their test's planned scope or assiduously meet the ensuing demands within a more comprehensive developmental process (such as that described in Downing & Haladyna, 2006).

Because reliability forms a foundation for validity, it is often among the first forms of evidence collected to support a test's use. Only the KSPT and VMPAC examined test–retest reliability, and only the VMPAC assessed interexaminer reliability. Neither effort met operational definitions for adequacy, however, due to failures to provide readily available information, such as the test–retest interval, statistical significance of reported reliability coefficients, and examiner characteristics.

Similar problems were identified with regard to evidence of validity. Despite operational definitions designed to provide great flexibility in how evidence was offered, most tests did not provide adequate evidence. Again, however, failure to meet the operational definitions was often the result of a failure to report information that was probably readily available to the test authors, or could have been, with adequate preparation and attention. For example, greater attention to selecting and describing participants and to reporting of statistical data would have led three tests to have produced satisfactory evidence in support of validity. Therefore, future test development efforts should incorporate more thorough descriptions of their methods.

At the outset, we anticipated that the quality of existing tests used for motor speech disorders in children was at risk because of the emerging knowledge base concerning these disorders, their changing nature over children's development, and difficulties associated with developing performance tests for young children. In fact, the deficiencies we observed appeared attributable to a lack of attention to basic psychometric standards (e.g., AERA et al., 1999). This oversight is understandable in that clinically oriented test authors are typically well trained in their content areas, but not

necessarily in the complex field of psychometrics or, especially, in the connection between test development and research (Streiner & Norman, 2003). The authors of these six tests should be commended for their efforts in providing much needed assessment tools. Their work will form the basis for further advances in tests of motor speech performance.

The pressures to produce measures that are short enough to be feasible for young test takers but long enough to suggest use for several purposes must surely be felt as part of the pressure to develop financially viable products. Because of the varying types of expertise and higher levels of funding required for an adequate program of research to offer evidential support for tests, collaborative and publicly funded efforts should be marshaled to help address test development in this area.

The future of test development in this area seems promising and worthy of broader collaborative efforts and more substantial financial support for at least two reasons. First, scholarly interest and activity related to test development is burgeoning. In addition to the traditional perspectives associated with psychological and educational testing (AERA et al., 1999; Downing & Haladyna, 2006), evidence-based practice perspectives associated with epidemiology and more recently with speech-language pathology (e.g., Dollaghan, 2004; Sackett, Richardson, Rosenberg, & Haynes, 2000; Spaulding et al., 2006; Straus, Richardson, Glasziou, & Haynes, 2005) provide ample fuel for future improvements. For older children (6 years and above), computerized efforts to address the motivational and measurement issues associated with maximal performance tasks seem promising (Rvachew et al., 2005).

Second, the knowledge base concerning motor speech disorders in children is growing in ways that can support test development, especially in the area of diagnosis. In particular, Shriberg and his colleagues are undertaking programmatic research designed to identify acoustic and genetic markers of CAS (e.g., Shriberg, 1993, 2003; Shriberg et al., 2006; Shriberg, Campbell, et al., 2003; Shriberg, Green, et al., 2003) These and related efforts, including those designed to identify correlated brain differences through imaging techniques, present existing and future test authors with measures that may eventually serve as gold standards. In the meantime, however, clinicians are in the position of having no tests that can be considered well developed for use with children with motor speech disorders. Within an evidence-based practice perspective, one is enjoined to find the best evidence (or test) available and to use it along with clinical experience and knowledge of the client (Sackett et al., 2000). Consequently, clinicians' knowledge of these disorders and clinical experience with them assume primary importance in determining the quality of decision making.

As part of their clinical decision making in children's motor speech disorders, clinicians may continue to turn to standardized tests, but they will want to do so with a philosophical appreciation of the current status of such measures. For example, they may choose to use one of the six standardized tests we reviewed, consider it a sample of behaviors that they believe are relevant to the decisions to be made, and report their findings with appropriate cautions. Alternatively, they may pursue the development of an informal measure themselves or adapt an existing tool. Any of these methods requires cautious use and interpretation because of their unknown or rudimentary claims to reliability and validity (McCauley, 2001; Vetter, 1988). Still, no alternative is as satisfying as having at least the option of choosing from an array of well-developed standardized tests. Therefore, clinicians will probably want to learn more about test development as a basis for urging publishers, funding agencies, researchers, and individual test authors to contribute to the development of better tests for children with motor speech disorders.

Acknowledgments

The authors wish to acknowledge David Ridge, Ruth Stoeckel, Barry Guitar, and Kimberly Bocian for their valuable contributions to this document. Dr. Strand's participation in this project was supported by a grant from The Mayo Clinic, Rochester, MN, CR-20.

References

Aase, D., Hovre, C., Krause, K., Schelfhout, S., Smith, J., & Carpenter, L. J. (2000). *Contextual Test of Articulation*. Eau Clair, WI: Thinking Publications.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *The standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Speech-Language-Hearing Association. (2007). *Childhood apraxia of speech* [Position statement]. Available from www.asha.org/policy.

Aronowitz, R. A. (2001). When do symptoms become a disease? *Annals of Internal Medicine, 134,* 803–808.

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

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Hillsdale, NJ: Erlbaum.

Blakeley, R. W. (2001). *Screening Test for Developmental Apraxia of Speech—Second Edition*. Austin, TX: Pro-Ed.

Buckendahl, C. W., & Plake, B. S. (2006). Evaluating tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 725–738). Mahwah, NJ: Erlbaum.

Buros Institute. (2006). Test Reviews Online. Retrieved February 11, 2006, from www.unl.edu/buros.

Caruso, A. J., & Strand, E. A. (1999). Motor speech disorders in children: Definitions, background, and a theoretical framework. In A. J. Caruso & E. A. Strand (Eds.), *Clinical management of motor speech disorders in children* (pp. 1–27). New York: Thieme.

Cizek, S. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225–260). Mahwah, NJ: Erlbaum.

Davis, B. L., Jakielski, K. J., & Marquardt, T. P. (1998). Developmental apraxia of speech: Determiners of differential diagnosis. *Clinical Linguistics and Phonetics, 12*(1), 25–45.

Davis, B. L., & Velleman, S. (2000). Differential diagnosis of developmental apraxia of speech in infants and toddlers. *Infant Toddler Intervention: The Transdisciplinary Journal, 10*(3), 177–192.

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

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders, 37,* 391–400.
Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Erlbaum.
Feinstein, A. R. (2001). The Blame-X syndrome: Problems and lessons in nosology, spectrum, and etiology. *Journal of Clinical Epidemiology, 4,* 433–439.
Fisher, H., & Logemann, J. (1971). *Fisher–Logemann Test of Articulation Competence*. Austin, TX: Pro-Ed.
Fletcher, R. W., & Fletcher, S. W. (2005). *Clinical epidemiology: The essentials* (4th ed.). Baltimore: Lippincott, Williams & Wilkins.
Forrest, K. (2003). Diagnostic criteria for developmental apraxia of speech used by clinical speech-language pathologists. *American Journal of Speech-Language Pathology, 12,* 376–380.
Fudala, J., & Reynolds, W. (2000). *Arizona Articulation Proficiency Scale, Third Edition*. Los Angeles: Western Psychological Services.
Goldman, R., & Fristoe, M. (2000). *Goldman Fristoe Test of Articulation—Second Edition*. Circle Pines, MN: AGS.
Guyette, T. W. (2001). Review of the Apraxia Profile. In B. S. Plake & J. C. Impara (Eds.), *The fourteenth mental measurements yearbook* (pp. 57–58). Lincoln, NE: Buros Institute of Mental Measurements.
Haladyna, T. M. (2006). Roles and importance of validity studies in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 739–755). Mahwah, NJ: Erlbaum.
Hayden, D., & Square, P. (1999). *Verbal Motor Production Assessment for Children*. San Antonio, TX: The Psychological Corporation.
Hickman, L. (1997). *Apraxia Profile*. San Antonio, TX: The Psychological Corporation.
Hodson, B. W. (2003). *Hodson Assessment of Phonological Patterns—Third Edition*. Austin, TX: Pro-Ed.
Jelm, J. M. (2001). *Verbal Dyspraxia Profile*. DeKalb, IL: Janelle.
Kaufman, N. (1995). *Kaufman Speech Praxis Test for Children*. Detroit, MI: Wayne State University Press.
Kent, R. D., Kent, J., & Rosenbek, J. (1987). Maximal performance tests of speech production. *Journal of Speech and Hearing Disorders, 52,* 367–387.
Khan, L. M., & Lewis, N. P. (2002). *Khan–Lewis Phonological Analysis—Second Edition*. Circle Pines, MN: AGS.
Lanphere, T. (1998). *Test of Articulation in Context*. Austin, TX: Pro-Ed.
Lewis, B. A., Freebairn, L. A., Hansen, A. J., Iyengar, S. K., & Taylor, H. G. (2004). School-age follow-up for children with childhood apraxia of speech. *Language, Speech, and Hearing Services in Schools, 35,* 122–140.
Linn, R. L. (2006). The standards for educational and psychological testing: Guidance in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 27–38). Mahwah, NJ: Erlbaum.
Masterson, J., & Bernhardt, B. (2002). *CAPES (Computerized Articulation and Phonological Evaluation)*. San Antonio, TX: The Psychological Corporation.
McCabe, P., Rosenthal, J. B., & McLeod, S. (1998). Features of developmental dyspraxia in the general speech-impaired population? *Clinical Linguistics and Phonetics, 12*(2), 105–126.
McCauley, R. J. (1996). Familiar strangers: Criterion-referenced measures in communication disorders. *Language, Speech, and Hearing Services in Schools, 27,* 122–131.
McCauley, R. J. (2001). *Assessment of language disorders in children*. Mahwah, NJ: Erlbaum.
McCauley, R. J. (2003). Review of Screening Test for Developmental Apraxia of Speech—Second Edition. In B. Plake, J. C. Impara, & R. A. Spies (Eds.), *The fifteenth mental measurements yearbook* (pp. 786–789). Austin, TX: Pro-Ed.
Merrell, A. W., & Plante, E. (1997). Norm-referenced test interpretation in the diagnostic test process. *Language, Speech, and Hearing Services in Schools, 28,* 50–58.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.
Messick, S. (1995). Validity of psychological assessment: Validity of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.
Pena, 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.
Pendergast, K., Dickey, S. E., Selmar, J. W., & Soder, A. L. (1997). *Photo Articulation Test, Third Edition*. San Antonio, TX: The Psychological Corporation.
Plake, B. S., & Impara, J. C. (Eds.). (2001). *The fourteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
Riski, J. E., & Witzel, M. A. (2001). *Oral Mechanism Exam for Children and Young Adults: Craniofacial and Oral Evaluation*. Circle Pines, MN: AGS.
Robbins, J., & Klee, T. (1987). Clinical assessment of oropharyngeal motor development in young children. *Journal of Speech and Hearing Research, 52,* 272–277.
Rvachew, S., Hodge, M., & Ohberg, A. (2005). Obtaining and interpreting maximum performance tasks from children: A tutorial. *Journal of Speech-Language Pathology and Audiology, 29*(4), 146–157.
Sackett, D. L., Richardson, W. S., Rosenberg, W. M. C., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM* (2nd ed.). London: Churchill-Livingstone.
Salvia, J., Ysseldyke, J., & (with Bolt, S.). (2007). *Assessment* (10th ed.). Boston: Houghton Mifflin.
Secord, W. (1981). *Test of Minimal Articulation Competence*. San Antonio, TX: The Psychological Corporation.
Secord, W., & Donohue, J. (1998). *Clinical Assessment of Articulation and Phonology.* Greenville, SC: Super Duper.
Shriberg, L. D. (1993). Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech and Hearing Research, 36,* 105–140.
Shriberg, L. D. (2003). Diagnostic markers for child speech-sound disorders: Introductory comments. *Clinical Linguistics and Phonetics, 17,* 501–505.
Shriberg, L. D., Aram, D. M., & Kwiatkowski, J. (1997a). Developmental apraxia of speech: I. Descriptive and theoretical perspectives. *Journal of Speech, Language, and Hearing Research, 40,* 273–285.
Shriberg, L. D., Aram, D. M., & Kwiatkowski, J. (1997b). Developmental apraxia of speech: III. A subtype marked by inappropriate stress. *Journal of Speech, Language, and Hearing Research, 40,* 313–337.
Shriberg, L. D., Ballard, K. J., Tomblin, J. B., Duffy, J. R., Odell, K. H., & Williams, C. A. (2006). Speech, prosody, and voice characteristics of a mother and daughter with a 7;13 translocation affecting FOXP2. *Journal of Speech, Language, and Hearing Research, 49,* 500–525.
Shriberg, L. D., Campbell, T. F., Karlsson, H. B., Brown, R. L., Mcsweeny, J. L., & Nadler, C. J. (2003). A diagnostic

marker for childhood apraxia of speech: The lexical stress ratio. *Clinical Linguistics and Phonetics, 17,* 549–574.

Shriberg, L. D., Green, J. R., Campbell, T. F., Mcsweeny, J. L., & Scheer, A. R. (2003). A diagnostic marker for childhood apraxia of speech: The coefficient of variation ratio. *Clinical Linguistics and Phonetics, 17,* 575–595.

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37,* 61–72.

St. Louis, K. O., & Ruscello, D. (2000). *Oral Speech Mechanism Screening Examination, Third Edition*. Austin, TX: Pro-Ed.

Strand, E. A. (2002). Childhood apraxia of speech: Suggested diagnostic markers for the younger child. In L. Shriberg & T. Campbell (Eds.), *Proceedings of the 2002 childhood apraxia of speech research symposium* (pp. 75–80). Carlsbad, CA: The Hendrix Foundation.

Strand, E. A., & McCauley, R. J. (1999). Assessment procedures for treatment planning in children with phonologic and motor speech disorders. In A. J. Caruso & E. A. Strand (Eds.), *Clinical management of motor speech disorders in children* (pp. 73–108). New York: Thieme.

Straus, S. E., Richardson, W. S., Glasziou, P., & Haynes, R. B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3rd ed.). New York: Elsevier/Churchill Livingston.

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use* (3rd ed.). New York: Oxford University Press.

Thoonen, G., Maassen, B., Wit, J., Gabreels, F., & Schreuder, R. (1996). The integrated use of maximum performance tasks in differential diagnosis evaluations among children with motor speech disorders. *Clinical Linguistics and Phonetics, 10,* 311–336.

Towne, R. L. (2001). Review of the Oral Speech Mechanism Screening Examination, Third Edition. In B. S. Plake & J. C. Impara (Eds.), *The fourteenth mental measurements yearbook* (pp. 868–869). Lincoln, NE: Buros Institute of Mental Measurements.

Vetter, D. (1988). Designing informal assessment procedures. In D. E. Yoder & R. D. Kent (Eds.), *Decision making in speech-language pathology* (pp. 192–193). Toronto, Ontario, Canada: Decker.

Weiss, C. E. (1980). *Weiss Comprehensive Articulation Test*. Austin, TX: Pro-Ed.

Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303–327). Mahwah, NJ: Erlbaum.

Wilcox, K., & Morris, S. (1999). *Children's Speech Intelligibility Measure*. San Antonio, TX: The Psychological Corporation.

Yorkston, K., Beukelman, D., Strand, E., & Bell, K. (1999). *Management of motor speech disorders in children and adults* (2nd ed.). Austin, TX: Pro-Ed.

Received August 19, 2005
Revision received December 22, 2006
Accepted July 25, 2007
DOI: 10.1044/1058-0360(2008/007)

Contact author: Rebecca McCauley, 402 Pomeroy Hall, University of Vermont, 489 Main Street, Burlington, VT 05405-0010. E-mail: rebecca.mccauley@uvm.edu.