Evaluating the Reliability and Validity of (Central) Auditory Processing Tests: A Preliminary Investigation

Jennifer C. Friberg, Ed.D.
Illinois State University
Normal, Illinois

Tena L. McNamara, Au.D.
Eastern Illinois University
Charleston, Illinois

Twenty-two criterion referenced and standardized tests commonly used to diagnose (central) auditory processing disorders were evaluated for both diagnostic accuracy and test validity. Tests were evaluated for evidence of diagnostic accuracy, level of acceptability of any identified diagnostic accuracy, and test validity for those tests with reported levels of diagnostic accuracy. Criteria for test validity were modified from McCauley and Swisher (1984) and McCauley (1996). Results indicated that 45% of reviewed tests had published evidence of diagnostic accuracy, although only 23% of tests met criteria for acceptable levels of both sensitivity and specificity. Evaluation of test validity indicated strengths in procedural aspects of test administration and weaknesses in various aspects of reliability and validity. Because sufficient evidence to support the reliability and validity of many (C)APD tests is not available in published data, findings indicated a clear need for educational audiologists to make (C)APD test selection decisions with care.

Introduction

The American Speech-Language-Hearing Association (ASHA) has prioritized evidence-based practice for all clinicians. As part of that process, ASHA has demanded that all diagnostic tools be evaluated for their ability to apply appraisal criteria (detailed in relevant research) for the identification of the most valid tools available for clinical use, particularly those used for the purposes of assessment and diagnosis (ASHA, 2005b).

Within the field of audiology, one area of clinical and diagnostic focus is particularly appropriate when considering test validity for the assessment of (central) auditory processing disorders ([C]APD). (Central) auditory processing disorders can be defined as difficulties in the processing of auditory information in the central nervous system (CNS), as demonstrated by poor performance in one or more of the following skills: sound localization and lateralization, auditory discrimination, auditory pattern recognition, temporal aspects of audition (including temporal integration, temporal discrimination, temporal ordering, and temporal masking), auditory performance in competing acoustic signals, and auditory performance with degraded acoustic signals (ASHA, 2005a. p. 2).

Historically, there has been significant controversy concerning the assessment of (C)APDs. Central auditory processing abilities are examined by audiologists using a combination of behavioral, electroacoustical, and electrophysiological approaches (ASHA, 2005a). Recognizing behavioral tests as being sensitive to lesions of the central auditory nervous system and important diagnostic data sources (due to the insight they provide into the functional listening abilities of clients), ASHA (2005a) has recommended that behavioral tests be used in conjunction with electroacoustical and electrophysiological measures to diagnose the presence of (C)APD. However, there is little consensus among professionals as to which tests should be utilized within the battery. While the use of electroacoustical and electrophysiological assessment techniques allows for the gathering of information relative to the neural function of the central auditory nervous system (Baran, 2007), electrophysiologic tests are not always readily available for diagnostic purposes. This is due to a lack of resources and equipment. Thus, many clinicians rely on electroacoustical and behavioral tests administered in clinically appropriate settings, as they are more readily available for diagnostic use (Emanuel, 2002; Jerger & Musiek, 2000).

With this in mind, clinicians must give careful consideration to which behavioral measures are most valid for clinical use when determining the presence or absence of (C)APD. With that said, there have been no studies conducted that review the validity of the various tests used in a battery to assess auditory processing abilities. This is despite a documented need for research focused on evaluating the validity and reliability of tests of central auditory function (Bellis, 2003; Keith, 2009a). Rather, several smaller studies have been conducted to look at isolated variables related to a test's validity for various individual (C)APD assessment

tools, including the Staggered Spondaic Word Test (SSW; Berrick, Shubow, Schultz, Freed, Fournier, & Hughes, 1984), low- and high-pass filtered versions of the Northwestern University Auditory Test No. 6 (NU-6; Bornstein, Wilson, & Cambron, 1994), Dichotic Digits Test (Kelly, 2007; Musiek, 1983b), Duration Pattern Test (Musiek, 1994), Frequency Pattern Test (Kelly, 2007; Musiek & Pinheiro, 1987b), Auditory Continuous Performance Test (Riccio, Cohen, Hynd, & Keith, 1996), Gaps-in-Noise Test (GIN; Shinn, Chermak, & Musiek, 2009), Random Gap Detection Test (Kelly, 2007), Compressed and Reverberated Words Test (Kelly, 2007), and the Masking Level Difference Test (MLD; Wilson, Moncrieff, Townsend, & Pillion, 2003).

Recognizing the paucity of research, Bellis (2003) indicated that "an understanding of issues surrounding validity and reliability of central auditory function tests is critical in order to determine the clinical utility of specific testing tools" and that the issue of test validity of (C)AP tests is an area of "further, much-needed research" (p. 202). Similarly, Keith (2009a) identified inadequacies within normative data provided for the vast majority of (C)APD assessment tools and indicated that this lack of normative data makes it extremely difficult to accurately judge whether or not a child exhibits a deficit in (C)AP abilities. Some tests' authors suggest that clinicians acquire and develop their own norms when using a particular (C)APD test in order to overcome this assessment challenge (Bellis, 1996; Emanuel, 2002; Kelly, 2007). However, acquiring one's own normative data can be problematic, due to subject variables, divergent testing procedures, and reduced quality control (Katz, Johnson, Brandner, Delagrange, Ferre, King, et al., 2002; Stewart & Kaminski, 2002). As collecting local normative data can be complicated, having well-established, representative, age-appropriate normative data provided by the test's author, test's publisher, or a researcher conducting large-scale research remains the ideal for (C)APD test administration (ASHA, 2006; Keith, 2009a).

It should be noted that within the field of communication sciences and disorders, most studies dealing with test validity have been conducted on tests relative to speech-language pathology. Predominantly, studies focusing on test validity in communication sciences and disorders use the specific psychometric criteria first used by McCauley and Swisher (1984) as the basis for evaluating standardized assessment tools. These criteria have become well established as acceptable and relevant to speech and language assessments. These authors would argue that the criteria are applicable for expansion into other areas of research within the communicative sciences and disorders, including the field of audiology. In fact, the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council for Measurement in Education (NCME) have collaborated to develop author guidelines for educational and psychological testing. These guidelines have become foundational to assessment practices for a broad range of specialists (AERA, 1999). Within these guidelines, the AERA, APA, and NCME clearly define behavioral tests as multidimensional tools that can be administered across a wide range of content areas to accomplish very similar ends (comparison of test subjects to a normative group or predetermined set of criteria; AERA, 1999). Specifically, these organizations recognize tests of cognitive processing, attention, auditory sensitivity, and "tests requiring reasoning and judgment as they relate to the processing and elaboration of complex sensory combinations and inputs" as having a psychological and/or educational basis (AERA, 1999, p. 123). Many different audiology assessment tools fit this description, including those used for diagnosing (C)APD.

McCauley and Swisher (1984) were among the first researchers in the field of communication sciences and disorders to assess the overall test validity of standardized assessment tools commonly used by practicing clinicians. As part of their research, McCauley and Swisher identified ten specific criteria related to the validity and reliability of standardized tests and applied them to over 30 different language assessment tools to judge the presence and/or absence of these criteria. These psychometric criteria were based on information provided on sample size, the normative sample, item analysis, measures of central tendency, concurrent validity, predictive validity, test/retest reliability, inter-examiner reliability, explanation of testing procedures, and testing qualifications. McCauley and Swisher hypothesized that tests with the most criteria met would be considered the most psychometrically valid of those reviewed and, thus, best for diagnostic use. Rather than identifying strong tests for diagnostic use by speech-language pathologists, this review highlighted the shortcomings of this cadre of tests. Results indicated that many of the assessment tools being used to identify speech and language disorders could not be used with any validity for such a purpose, as only 12 of the 30 assessment tools reviewed met even three of the original ten criteria. The psychometric criteria applied in this study have become the hallmark for assessing test validity to this point in the field of communication sciences and disorders, as they have been used repeatedly for this purpose for the last two decades (Friberg, 2010; Hutchinson, 1996; Mikucki & Larrivee, 2006; Plante & Vance, 1994). Taking the lead from this foundational study (McCauley & Swisher, 1984), several other researchers have conducted in-depth studies of commonly used assessment tools focusing on different aspects of test validity within the fields of audiology and speech/language pathology. These researchers have identified issues related to the composition of normative samples (Pena, Spaulding, & Plante, 2006), use of standardized assessment tools with clients

from culturally and linguistically diverse backgrounds (Restrepo & Silverman, 2001; Thomas-Tate, Washington, & Edwards, 2004; Yavas & Goldstein, 1998), and interpretation and application of standardized assessment scores (Pena et al., 2006; Plante & Vance, 1994, 1995). In each of these studies, some threat was found to the test validity of the instrument being used to evaluate an individual's communication skills, indicating that the overall validity associated with scores obtained on many clinically applied behavioral testing instruments remains an area of concern.

Recently, the notion of diagnostic accuracy has been featured in research related to an assessment tool's test validity (Friberg, 2010; Spaulding, Plante, & Farinella, 2006). This shift represents a new direction for researchers studying test validity in communication sciences and disorders because earlier research did not address this notion directly. Diagnostic accuracy refers to the degree with which a given assessment tool is able to diagnose the presence or absence of a disorder accurately. Diagnostic accuracy is of particular concern, as it has been determined that even tests with acceptable levels of test validity cannot always discriminate disordered skills from those considered to be more reflective of typically developing children (Gray, Plante, Vance & Henrichsen, 1999; Plante & Vance, 1994, 1995). To this end, it has been suggested that it might be of greater importance to evaluate the diagnostic accuracy of a particular assessment tool than it is to focus on other psychometric criteria used in the past. The diagnostic accuracy of a test indicates its overall accuracy of diagnosis, and ignoring this variable could lead to the improper identification of a disorder, with children being identified as disordered when they actually have typical functioning or, conversely, identified as being typically developing when a disorder is actually present (Dollaghan, 2004; Spaulding et al., 2006). Consequently, it has been suggested that it is inappropriate to assess the overall test validity of assessment tools for which data related to the diagnostic accuracy of the tests are not reported (Spaulding et al., 2006).

Diagnostic accuracy of a particular assessment tool is often described using two different measures: sensitivity and specificity. Sensitivity is the likelihood that a child who has previously been diagnosed as disordered is found to be disordered when using a different (but related) assessment tool. Conversely, specificity is the likelihood that a child considered to be typically developing in the area being assessed is identified as such, again using a different, but related, test (Dollaghan, 2004; Spaulding et al., 2006). Possessing this information related to the diagnostic accuracy of any particular assessment tool is critical, as lower than ideal levels of either sensitivity or specificity can lead to misdiagnosis, inaccurate eligibility determination, and possibly the provision of inappropriate services. Within the standardization process for assessment tools, sensitivity and specificity are measured by percent, with values indicating the overall accuracy of a particular assessment tool to make a valid diagnosis. Because it is important to ensure diagnostic accuracy, Plante and Vance (1994) and Dollaghan (2004) suggest that the threshold values for acceptable levels of sensitivity and specificity should be 80% or greater (Plante & Vance 1994), indicating that at least 80% of the time, children are correctly diagnosed when a particular assessment battery is administered. Without acceptable levels of sensitivity and specificity, results collected from an assessment tool cannot be considered a valid measure of the child's performance.

A great deal of information on the sensitivity of tests used in the assessment of (C)APD has been based on the performance of adults who have verified lesions in the auditory cortex. For example, a deficit in frequency pattern recognition has been observed in adults with a compromised central auditory system (Musiek & Pinheiro, 1987b). From this data on adults with a damaged auditory cortex, interest arose concerning the prospect of using these behavioral tests in children to diagnose (C)APDs. However, while it is well understood that performance on tests for adults cannot be compared equally to the performance of children, securing children to serve within a normative sample who have no known auditory lesions is not an easy task to accomplish. Thus, in order to accurately diagnose (C)APD in children, weaknesses on behavioral tests must be connected with difficulties that children are experiencing in the classroom. With the understanding that performance on behavioral tests constitutes an important piece of the (C)APD diagnostic puzzle, clinicians must be able to trust the results obtained using these measures. Thus, the presence of acceptable levels of diagnostic accuracy for all behavioral assessment tools does remain the gold standard for accurate (C)APD diagnosis.

Purpose of Research

Research has repeatedly demonstrated that clinicians cannot discount the need to identify strengths and weaknesses relative to a test's validity prior to its use. Such oversight might well lead to inaccurate diagnostic decision making. Additionally, research has suggested that paramount to the notion of test validity, diagnostic accuracy must be a central consideration in the selection of tests meant to diagnose the presence or absence of a particular disorder. Therefore, the current research sought to examine the data provided within examiner's manuals and peer-reviewed, published research for a variety of behavioral (C)APD assessment tools to accomplish the following: 1) identify (C)APD tests that report information related to their diagnostic accuracy, 2) evaluate the acceptability of any diagnostic accuracy evidence found, and 3) assess the test validity and reliability of those (C)APD tests found to have reported levels of diagnostic accuracy.

Method

Selection of Assessment Tools for Evaluation

The purpose of this study was to examine the diagnostic accuracy and test validity of commercially available behavioral tests used to diagnose the presence or absence of (C)APD. Determining which tests to include in this study was complicated in light of the fact that no standard protocols exist for determining which tests should be administered for the most efficient, accurate diagnosis of (C)APD (Emanuel, 2002; Singer, Hurley, & Preece, 1998). In an effort to comprehensively review all relevant (C)APD assessment tools, behavioral tests identified by Chermak et al. (2007) and Emanuel (2002) as being frequently utilized by practicing clinicians were reviewed for this study with few exceptions. Emanuel (2002) found the Rapid Alternating Speech Perception Test (RASP; Willeford, 1976) was commonly administered by clinicians; however, due to recommendations that the RASP not be used secondary to poor quality recordings and norms (Shea & Raffin, 1983), this test was not considered in the present review. Three newly published, standardized (C)AP tests were added for review: the Multiple Auditory Processing Assessment (MAPA; Schow, Seikel, Brockett, & Whitaker, 2007), the SCAN-3 for Adolescents and Adults (SCAN-3:A; Keith, 2009b) and the SCAN-3 for Children (SCAN-3:C; Keith, 2009c). Based on these guidelines, 22 assessments commonly employed by audiologists for (C)APD testing were secured for review. These 22 tests were categorized as either standardized or criterion-referenced, based on the presence or absence of a normative sample.

Fifteen of the 22 selected tests were found to be criterion-referenced. These tests included: the Bamford-Kowal-Bench Speech In Noise Test (BKB-SIN; Etymotic Research, Inc., 2005), Competing Sentences (CS; Willeford, 1977), Dichotic Digits Test (DDT; Auditec, n.d.), Dichotic Digits Test (DDT; Musiek, 1983b), Duration Pattern Sequence Test (DPST; Auditec, n.d.), Duration Pattern Test (DPT; Musiek, 1994; Musiek, Baran, & Pinheiro, 1990), Frequency Pattern Test (FPT; Musiek & Pinheiro, 1987b), Gaps-in-Noise Test (GIN; Musiek, 2005), Low-Pass Filtered NU-6 Test (LPF; Auditec, n.d.), Masking Level Difference Test (MLD; Auditec, 2003), Pitch Pattern Sequence Test-Adult Version (PPS-A; Pinheiro, 1977.), Pitch Pattern Sequence Test-Child Version (PPS-C; Pinheiro, 1977), Quick Speech-in-Noise Test (QuickSIN; Etymotic Research, 2001), Random Gap Detection Test (RGDT; Keith, 2000), and Spondaic Binaural Fusion Test (SBF; Auditec, n.d.).

Seven assessment tools were found as standardized and were selected for review. Four of these tests were identified as being amongst the most commonly administered standardized tests for (C)APD (Emanuel, 2002; Chermak et al., 2007): the Auditory Continuous Performance Test (ACPT; Keith, 1994), Auditory Fusion Test-Revised (AFT-R; Keith & McCrosky, 1996), Selective Auditory Attention Test (SAAT; Cherry, 1998), and the Staggered Spondaic Word Test, 5th Edition (SSW-5; Katz, 1998). As previously stated, three recently published standardized tests were added to those identified as being commonly administered: the MAPA (Schow, Seikel, Brockett, & Whitaker, 2007), the SCAN-3:A (Keith, 2009b) and the SCAN-3:C (Keith, 2009c). While not specifically identified as being commonly administered in their revised form, earlier forms of the SCAN-3:A and SCAN-3:C were identified as commonly administered by clinicians (Emanuel, 2002).

Data Collection and Analysis

A similar process was utilized to review (C)APD assessment tools as was used in previous studies concerning tests from other areas of communication sciences and disorders (Friberg, 2010; McCauley & Swisher, 1984; Plante & Vance, 1994). Initially, two communication sciences and disorders graduate students reviewed each of the assessment tools used in this study to judge the presence or absence of information related to each test's diagnostic accuracy, then to selected psychometric criteria, as appropriate. Both students received a training session with the first author of this study and were provided with guidelines to judge whether tests demonstrated specific criterion at an acceptable level. Trainings for student data collectors were two hours in duration and focused on identification of selected psychometric criteria in sample tests from a related professional field (speech-language pathology). After this initial training session, students were encouraged to contact the authors of this study to resolve any questions that arose in the data collection process.

In completing these analyses, graduate student reviewers consulted the examiner's manual provided as part of each commercially purchased assessment tool studied. The examiner's manuals were critically evaluated to determine whether information was available to indicate the presence or absence of criteria selected for use in this study. Those criteria judged to be present were marked as (+), while those criteria judged to be absent were marked as (-). Immediately, it was evident that few of the test's examiner's manuals contained information relative to the criteria being analyzed because few test manuals contained any evidence of validity as measured by the criteria used in this study. Thus, the authors of this study undertook an extensive literature search to document source data available within original research papers relative to the test validity of (C)APD assessment tools. Searches using the name of each test reviewed and keywords (e.g., [C]APD diagnostic, test validity, diagnostic accuracy, efficiency) were conducted, yielding several studies that could be used for analyses. References cited by authors of the tests were also obtained, which included test protocols and measurements. Information collected

in this literature review process was considered in the resulting analysis. Credit was assigned if sufficient evidence indicated the presence or absence of each criterion within a well-executed original research study.

At the conclusion of data collection, all results (from examiner's manuals and from published source data) were pooled and compared for agreement. Initial review of data indicated 97% agreement across raters, as the presence and or absence of each criterion was fairly clear to distinguish for the tests evaluated. Discrepancies (disagreements) amongst data were only found on one criterion (criterion #6, below) from the evaluation of standardized assessments dealing with item analysis. All discrepancies were addressed following procedures reflective of those used in previous research (Friberg, 2010; McCauley & Swisher, 1984). That is, the examiner's manual was revisited by the first author and the student data collector(s) to resolve any disparity in ratings. As a result of this process, three ratings were modified for reporting.

Diagnostic accuracy criteria. Following the initial evaluation of the 22 tests selected for use in this study, those tests found to have no evidence of diagnostic accuracy were reported as such and eliminated from further review. This procedure was consistent with the recommendations of Spaulding et al. (2006), who suggested the inappropriateness of assessing the overall test validity of assessment tools for which data related to their diagnostic accuracy are not reported. Therefore, the tests in which evidence of diagnostic accuracy were identified were further evaluated to ascertain whether each test possessed an "acceptable" level of diagnostic accuracy for clinical use. That is, levels of sensitivity and specificity had to be .80 or greater, in accordance with the recommendations of Plante and Vance (1994) and Dollaghan (2004).

Psychometric validity criteria. Each of the assessment tools with reported levels of diagnostic accuracy were subsequently rated for the presence or absence of specific criteria related to their validity. It was necessary to employ the use of different types of psychometric criteria based on whether an assessment tool was categorized as criterion-referenced or standardized. The following section describes the different criteria utilized to review each of the selected assessment tools.

Criterion-referenced tests. McCauley (1996) described guidelines for developers and users of criterion-referenced tests, which identified strengths and weaknesses relative to a given test's design and structure. Specifically, six guidelines related to a test's overall test validity were presented, with suggestions of how test users could look for certain types of evidence to support or refute a particular test's clinical use. These were not hard and fast recommendations, but rather suggestions for clinicians to consider when evaluating diagnostic tests. These guidelines were used to form the foundation from which criterion-referenced tests were evaluated within this current study. To evaluate the qualities of the criterion-referenced tests found to have evidence of diagnostic accuracy, the authors of the current study carefully considered each of McCauley's guidelines and determined what evidence tests would need in order to possess acceptable test validity. Each item of evidence needed by a test was established as a separate criterion, and each criterion-referenced test was evaluated for the presence and/or absence of these separate criteria. Three of McCauley's guidelines were used as specific individual criterion to judge the overall test validity of

Table 1. **Psychometric Criteria used in the Evaluation of Criterion-Referenced Tests**

Criteria #	Description of Criteria[a]
1	Clear definition of test domain, with inclusion of the following information:
	a. Clear definition of behavior assessed
	b. Statement of tasks to be completed
	c. Plan guiding item construction/item analysis
2	Evidence of validity, with inclusion of data reflecting the test's:
	a. Specific criteria for pass/fail scoring
	b. Predictive validity
	c. Concurrent validity
	d. Sensitivity of test
	e. Specificity of test
3	Evidence of reliability, with inclusion of data reflecting the test's:
	a. Inter-rater reliability
	b. Test/re-test reliability
4	Careful description of test takers used in studies of reliability/validity
5	Detailed description of test administration
6	Detailed description of user qualifications

[a] Description of psychometric criteria used for evaluating the test validity of criterion-referenced (C)APD tools is available in McCauley (1996).

each separate assessment tool. The remaining three guidelines were expanded, and in doing so, seven sub-criteria were created. Therefore, a total of eleven distinct criteria were applied to each of the criterion-referenced tests. Table 1 lists each of the criteria used in assessing criterion-referenced tests as part of this current study. These criteria are briefly explained in the following section:

- ***Criterion 1*:** Information should be provided that allows for a clear description of the test's overall scope and structure. Tests were classified as meeting Criterion 1 if information was available related to the following subcriteria: Criterion 1a (provision of a clear definition of the test's purpose/ behaviors), Criterion 1b (inclusion of a statement of tasks to be completed as part of the test), and Criterion 1c (provision of a plan guiding item construction/item analysis).
- ***Criterion 2:*** Data should be available to provide evidence of a test's validity prior to its diagnostic use. Tests were classified as meeting Criterion 2 if information was available related to the following subcriteria: Criterion 2a (presence of specific criteria for pass/fail scoring), Criterion 2b (evidence of predictive validity), and Criterion 2c (evidence of concurrent validity).
- ***Criterion 3:*** The consistency and stability of scores obtained on a test over time in various testing situations must be established. Tests were classified as meeting Criterion 3 if information was available related to the following subcriteria: Criterion 3a (test should have a reported inter-rater reliability coefficient of .90 or greater) and Criterion 3b (test should have a test-retest reliability coefficient of .90 or greater).
- ***Criterion 4:*** Each test must provide a thorough description of the test takers who participated in studies of the test's validity. Tests were classified as meeting Criterion 4 if the test provided information for clinicians to review regarding gender, age, grade, socio-economic status, impairment status, and/or geographic distribution of the validation sample.
- ***Criterion 5:*** Each test must provide a clear and detailed description of test administration. Tests were classified as meeting Criterion 5 if the manual provided instructions detailed enough for standard, straightforward administration by a qualified clinician in a manner that is "in compliance with recommended procedures" to "increase the likelihood that the measure will function as intended" (McCauley, 1996, p. 128).
- ***Criterion 6:*** Tests must provide a clear description of the requirements for an examiner to be deemed qualified to administer the test in question. Tests were classified as meeting Criterion 6 if the examiner's manual elaborated on the educational training needed to administer and interpret the results of the test.

Standardized assessment tools. With few modifications, criteria developed by McCauley and Swisher (1984) have been utilized for judging validity of the selected standardized (C)APD assessment tools. Beyond these original ten criteria, one new criterion was added, and one existing criterion was modified. These changes were made to identify threats to test validity more completely, relative to suggestions in recent research advocating for a broadening of focus when considering the test validity of standardized tests. Criteria that remained unchanged from McCauley and Swisher's prior work (1984) are identified and briefly described in Table 2. New and modified psychometric criteria used to review assessment tools in this study are described briefly below.

Clearly defined standardization sample. This criterion was part of McCauley and Swisher's original research (1984), and required that all standardization samples contain information that would allow clinicians to consider the characteristics of the normative sample to ensure that a test that might be administered is representative of the child(ren) to be tested. The importance of a clearly defined standardization sample cannot be understated because test scores obtained from a standardized test (using norms gathered from individuals not reflecting the demographics of the child being tested) cannot be considered a valid representation of strengths or weaknesses with regard to the skill being tested. Furthermore, researchers have stated that "the most compelling [diagnostic] evidence" is found when standardization samples are "broadly representative of the range of individuals about whom the diagnostic decision is to be made" (Dollaghan, 2004, p. 395-6). Thus, a clearly defined standardization sample gives clinicians knowledge of what subgroups of individuals to whom their client(s) will be compared, which will help to inform test selection and ensure a more valid measurement outcome.

Originally, there were only three demographic categories listed as part of this criterion: geographic representation, socioeconomic status, and the communication status of those in the normative group (typical vs. atypical skills; McCauley & Swisher, 1984). These three demographic categories have been expanded to further consider the normative sample for each assessment tool evaluated in this study. Specifically, Spaulding et al. (2006) suggested that any consideration of the normative sample should include the addition of age and gender distribution, as well as ethnic background. Additionally, Entwisle and Astone (1994) indicated that parental education level could serve as an acceptable measure of socio-economic status, as these variables have been found to correlate with one another. Thus, a test was considered to have met this criterion if it provided information about its standardization sample related to geographic representation, socio-economic status representation, gender distribution, ethnic representation, sample +/- impairment(s), and age distribution.

Purpose of assessment tool. Identification of the purpose of an assessment tool has recently been emphasized for a variety of important reasons (Merrell & Plante, 1997; Plante & Vance, 1995; Spaulding et al., 2006). Specific to standardized assessment tools, standardized tests are often administered to document the existence (or non-existence) of a disorder. Similarly, a test might be given in order to quantify the severity of an existing disorder.

A lack of information detailing the purpose of a given assessment tool may compromise the validity of the data collected when the test is administered. Clinicians might make decisions related to eligibility and treatment based on results for a test meant to be used for one purpose that was actually administered to serve an entirely different clinical function. Thus, information related to a test's purpose(s) is a critical component for any assessment tool.

For the purposes of this current review, an assessment tool was considered to have provided an acceptable amount of information related to the purpose of the test if the examiner's manual contained a section outlining the specific purpose(s) of the test in question.

Results

Diagnostic Accuracy

Criterion-referenced tests. Of the 15 criterion-referenced tests reviewed, five tests (CS, Willeford, 1977; DDT, Auditec, n.d.; DPT, Musiek, 1994; FPT, Musiek & Pinheiro, 1987b; GIN, Musiek, 2005) were found to have reported levels of sensitivity and specificity for clinical review. One test was found to have reported levels of sensitivity, but no reported specificity (DD, Musiek, 1983b) and nine other tests (BKB-SIN, Etymotic Research, Inc., 2005; DPST, Auditec, n.d.; LPF, Auditec, n.d.; MLD, Auditec, n.d.; PPS-A, Pinheiro, 1977; PPS-C, Pinheiro, 1977; QuickSIN; Etymotic Research, 2001; RGDT, Keith, 2000; SBF, Auditec, n.d.) were found to have no reported evidence of either sensitivity or specificity in their examiner's manuals or peer-reviewed research publications.

With regard to the level of diagnostic accuracy identified, only the DPT (Musiek, 1994) and FPT (Musiek & Pinheiro, 1987b) were found to have acceptable levels of diagnostic accuracy, as

Table 2. **Summary of McCauley and Swisher (1984) Psychometric Criteria Used For Analysis of (C)APD Assessments**

Psychometric Criteria	Description of Criteria[a]
Adequate Sample Size	Must have at least 100 participants in each comparison subgroup within the normative sample
Evidence of Item Analysis	Test items for the test in question are scrutinized to ensure that they test what they purport to measure.
Measures of Central Tendency Reported	Means and standard deviations of normative sample must be available to allow for flexibility in comparing scores/data within the test in question.
Concurrent Validity is Documented	Results from another, similar, standardized instrument agree with the results obtained from the test in question.
Predictive Validity is Documented	Performance on the test in question is predictive of performance on other, less formal measure (observation, etc.) in a more functional setting.
Test/Retest Reliability is Reported	Ensures that test scores on the test in question are stable over time (correlation of .90 or greater must be reported)
Inter-Examiner Reliability is Reported	Ensures that test scores on the test in question don't fluctuate depending if different clinicians administer the test (correlation of .90 or greater must be reported)
Testing Procedures Explained Sufficiently	Sufficient detail must be provided to ensure that the test can be administered in a way that mirrors test administration for normative sample.
Testing Qualifications Explicitly Stated	Special training/qualifications for test administrators must be clearly stated.

[a] Detailed descriptions of psychometric criteria are available in McCauley and Swisher (1984).

both possessed sensitivity and specificity values of .80 or greater (Dollaghan, 2004). It should be noted that three criterion-referenced tests were reported to have acceptable levels of specificity, but were lacking acceptable levels of sensitivity (CS, Wellford, 1977; DDT, Auditec, n.d.; and GIN, Musiek, 2005). Only one test was reported to have an acceptable level of sensitivity, though no data was available to judge the specificity of the test (DD; Musiek, 1983b).

Standardized tests. Of the seven standardized tests reviewed, three tests (MAPA, Schow et al., 2007; SCAN-3:A, Keith, 2009b; SCAN-3:C, Keith, 2009c) were each found to have reported levels of sensitivity and specificity above .80, indicating acceptable levels of diagnostic accuracy (Dollaghan, 2004). It should be noted that the sensitivity and specificity for the SCAN-3:A and SCAN-3:C were acceptable only at specific cut scores, specified within each test's examiner's manual.

One test (SSW-5; Katz, 1998) was found to have reported levels of specificity, but not sensitivity, and three tests were found to have no reported evidence of diagnostic accuracy published at all (ACPT, Keith, 1994; AFT-R, Keith & McCrosky, 1996; SAAT, Cherry, 1998). Table 3 contains a listing of the tests found to have evidence of diagnostic accuracy, the reported levels of sensitivity and specificity, and the source from which these data were found.

Evidence of Test Validity

A complete accounting of the presence and absence of each selected psychometric criterion can be found in Table 4 (Criterion-Referenced Tests) and Table 5 (Standardized Assessment Tools). The criteria are arranged by assessment tool.

Criterion-referenced tests. Of the six criterion-referenced tests evaluated to determine their level of test validity, no assessment tool was able to meet all 11 criteria applied to them as part of this study. Rather, the six criterion-referenced assessment tools were found to possess a range from three to six of the evaluated criteria, with four tests meeting three criteria (CS, Willeford, 1977; DDT, Auditec, n.d.; DPT, Musiek, 1994; and FPT, Musiek & Pinheiro, 1987b) and one test (GIN, Musiek, 2005) meeting six criteria.

Table 3. **Evidence of Diagnostic Accuracy for (C)APD Assessment Tools**

Name of (C)APD Assessment Tool	Data Source	Level of Sensitivity	Level of Specificity
Competing Sentences (CS)	Schow et al. (2007)	.25	1.0
Dichotic Digits Test (DDT)	Auditec (n.d.)	.30	1.0
Dichotic Digits Test (DD)	Musiek (1983)	.81	Not reported
Duration Pattern Test (DPT)	Musiek (1994); Musiek et al. (1990)	.86	.92
Frequency Pattern Test (FTP)	Musiek and Pinheiro (1987)	.80	.88
Gaps-in-Noise Test (GIN)	Shinn et al. (2009)	.67	.94
Multiple Auditory Processing Assessment (MAPA)	Shiffman (1999)	.83	.85
SCAN-3 for Adolescents and Adults (SCAN-3:A)	Keith, 2009b	.93	.85
SCAN-3 for Children (SCAN-3C)	Keith, 2009c	.90	.83
Staggered Spondaic Word Test, 5th Edition	Berrick et al. (1984)	Not reported	.77

Note: Abbreviations were used to denote the following (C)APD tests: Competing Sentences (CS; Willeford, 1977), Dichotic Digits Test (DDT; Auditec, n.d.), Dichotic Digits Test (DD; Musiek, 1983b), Duration Pattern Test (DPT; Musiek, 1994; Musiek, Baran, & Pinheiro,1990), Frequency Pattern Test (FPT; Musiek & Pinheiro, 1987), Gaps-in-Noise Test (GIN; Musiek, 2005), Multiple Auditory Processing Assessment (MAPA; Schow, Seikel, Brockett, & Whitaker, 2007), SCAN-3 for Adolescents and Adults (SCAN-3:A; Keith, 2009b), SCAN-3 for Children (SCAN-3:C; Keith, 2009c), and the Staggered Spondaic Word Test, 5th Edition (SSW-5; Katz, 1998).

Of the criteria selected for use in this study, criterion-referenced tests were most able to meet standards relative to detailed description of test administration, clear definition of test domain, and careful description of test takers used in studies of reliability and validity. All criterion-referenced tests provided specific criteria for pass/fail scoring. These tests were less successful at meeting criteria relative to evidence of validity, evidence of reliability, and detailed description of user qualifications. None of the criterion-referenced tests provided information relative to predictive validity, concurrent validity, and inter-rater reliability. The majority of data reviewed with regard to the criterion-referenced tests were found within individual test examiner's manuals with the exception of data related to test sensitivity, which was most often reported in published literature (Musiek, 1983a; Musiek, Baran & Pinheiro, 1990; Musiek & Pinheiro, 1987a; Shinn, Chermak, & Musiek, 2009).

Standardized assessment tools. Of the four standardized assessment tools evaluated for test validity within this study, none met all 11 criteria applied to them. Rather, the SCAN-3:A (Keith, 2009b) and SCAN-3:C (Keith, 2009c) each met seven criteria, the MAPA (Schow et al., 2007) met six criteria, and the SSW-5 (Katz, 1998) met five criteria. Of the criteria applied to the standardized tests analyzed in this study, all tests met four criteria uniformly: identification of test purpose, adequate explanation of testing procedures, evidence of item analysis, and reporting of measures of central tendency. Conversely, no test reported data relative to concurrent or predictive validity. Only the SCAN-3:A and SCAN-3:C provided information to clearly describe the normative sample used to standardize the test for clinical use. All data reported for standardized (C)APD assessment tools were found in the test examiner's manuals, with the exception of information relative to the SSW-5, which was found in published research (Berrick et al., 1984).

Table 4. **Results from Psychometric Analysis of Criterion-Referenced (C)APD Tests**

Criteria Descriptors		CS	DDT	DD	DPT	FPT	GIN
1	Clear Definition of Test Domain						
a	Clear definition of behavior assessed	-	-	+	-	-	+
b	Statement of tasks to be completed	+	+	+	+	-	+
c	Plan guiding item construction/item analysis	-	-	-	-	-	-
2	Evidence of Validity						
a	Specific criteria for pass/fail scoring	+	+	+	+	+	+
b	Predictive validity	-	-	-	-	-	-
c	Concurrent validity	-	-	-	-	-	-
3	Evidence of Reliability						
a	Inter-rater reliability	-	-	-	-	-	-
b	Test/retest reliability	-	-	-[a]	-	-	+[b]
4	Careful description of test takers used in studies of reliability and validity	-	-	-	-	+	+
5	Detailed description of test administration	+	+	+	+	+	+
6	Detailed description of user qualifications	-	-	-	-	-	-
	Total Criteria Met:	3/11	3/11	4/11	3/11	3/11	6/11

Note: Abbreviations were used above to denote the following criterion-referenced tests: Competing Sentences (CS; Willeford, 1977), Dichotic Digits Test (DDT; Auditec, n.d.), Dichotic Digits Test (DD; Musiek, 1983b), Duration Pattern Test (DPT; Musiek, 1994; Musiek, Baran, & Pinheiro, 1990), Frequency Pattern Test (FPT; Musiek & Pinheiro, 1987b), Gaps-in-Noise Test (GIN; Musiek, 2005)

[a] Found in Musiek, Gollegly, Kibbe, and Verkest-Lenz (1991); [b] Found in Shinn, Chermak, and Musiek (2009)

Discussion

Many audiologists rely on behavioral tests to diagnose (C)APD. Behavioral tests have offered audiologists a fairly inexpensive and readily obtainable means for assessing auditory processing skills, particularly in children (Emanuel, 2002; Jerger & Musiek, 2000). Information from behavioral tests can also provide an understanding of the auditory tasks on which a person may have the most difficulty, as these measures may carry significant meaning for how an individual performs in an everyday listening environment. For these reasons, it is important that clinicians have the opportunity to utilize tests that have acceptable levels of diagnostic accuracy and adequate test validity. The current review of commonly utilized (C)APD assessment tools, both standardized and criterion-referenced, highlights many important considerations for clinicians engaging in such diagnostic endeavors.

It should be noted that for several tests that were analyzed, tests' authors made reference to published literature that offered additional diagnostic accuracy and general psychometric data beyond that provided within the examiner's manuals. In an effort to be comprehensive in this review of (C)APD assessment tools, this information was obtained and included in the study. The question remains, however, whether data reflective of measures of each test's overall validity belongs within the examiner's manual or within outside, refereed research published in relevant journals. McCauley (1996) urges that

examiner's manuals need to specify not only an accounting of tasks to be accomplished during the administration of a test, but a plan for item construction and analysis, as well. Without this information provided, it is possible that during test administration, a clinician might "misconstrue the test takers' competencies," therefore undermining the usefulness of any results collected within a given test (McCauley, 1996, p. 126). Tests evaluated for this study were varied in their success in including these categories of information within their examiner's manuals. The majority of standardized and criterion-referenced assessment tools provided a clear statement of tasks to be accomplished, yet only the standardized tests evaluated here included information relative to item analysis for consideration. None of the criterion-referenced tests offered this information for diagnosticians. It would seem that, at a minimum, this might be a place to start with revisions of existing (C)APD tests and the introduction of new tests in this area. Publishing the remainder of psychometric data in peer-refereed journals could be deemed as acceptable, so long as clinicians were able to easily identify and secure such research for their review prior to selecting a (C)APD test for administration. It should be noted that Emanuel (2002) found commercially available tests that included detailed administration, interpretation, and scoring procedures as part of the testing package were most widely used for diagnosing (C)APD.

Overall, it would seem that none of the tests evaluated within this study exemplifies a precise diagnostic gold standard for (C)APD, as no test possessed all criteria applied to assess either diagnostic accuracy or test validity. Results from this study highlighted the strengths and weaknesses inherent within each of the tests analyzed. These strengths and weaknesses must be identified in order to inform selection of valid tests for clinical diagnosis of (C)APD. Clinical audiologists have the onus of utilizing evidence-based practices to select diagnostic tools (ASHA, 2005b), yet it is a complicated task to interpret research findings to compare and contrast assessment tools for clinical use. Findings from this research can serve as a starting point for such diagnostic decision making. Knowing the diagnostic accuracy and test validity of assessment tools available for use constitutes the first step in mitigating the threat of clinical misdiagnosis (Plante & Vance, 1994).

Table 5. **Results from Psychometric Analysis of Standardized (C)APD Assessments**

	Criteria Description	MAPA	SSW-5	SCAN-3:A	SCAN-3:C
1	Test Purpose Identified	+	+	+	+
2	Tester Qualifications	+	-	+	+
3	Procedures Explained	+	+	+	+
4	Adequate Sample Size	+	+	-	-
5	Sample Clearly Defined				
	a. geographic representation	-	-	+	+
	b. parent education/SES	-	-	+	+
	c. gender distribution	-	+	+	+
	d. ethnic representation	-	-	+	+
	e. +/- impairment	-	+[a]	+	+
	f. age distribution	+	+	+	+
6	Evidence of Item Analysis	+	+	+	+
7	Measures of Central Tendency	+	+[a]	+	+
8	Concurrent Validity	-	-	-	-
9	Predictive Validity	-	-	-	-
10	Test/Retest Reliability (>.90)	-	-	-	-
11	Inter-Examiner Reliability (>.90)	-	-	+	+
	# Criteria Met (per assessment tool)	6/11	5/11	7/11	7/11

Note: Abbreviations were used above to denote the following standardized tests: Multiple Auditory Processing Assessment (MAPA; Schow, Seikel, Brockett, & Whitaker, 2007), SCAN-3 for Adolescents and Adults (SCAN-3:A; Keith, 2009b), SCAN-3 for Children (SCAN-3:C; Keith, 2009c), and the Staggered Spondaic Word Test, 5th Edition (SSW-5; Katz, 1998).

[a] Found in Berrick, Shubow, Schultz, Freed, Founier, and Hughes (1984).

Data collected were reflective of several key ideas and trends across all assessment tools analyzed. Most notably, less than half of the tests analyzed for this study were found to have any published data reflecting their diagnostic accuracy. ASHA demands that clinicians use evidence-based techniques to diagnose (C)APDs (2005a), and the use of tools that accurately identify the presence or absence of any disorder is fundamental for clinical practice based on research and evidence to support it (McCauley & Swisher, 1984; Plante & Vance, 1994; Spauding et al., 2006).

Looking specifically at test validity, the majority of assessment tools evaluated in this study have acceptable levels of information provided relative to the more procedural aspects of test administration (e.g., clear definition of test domain, detailed description of test administration), yet they lack supporting data relative to the foundational constructs of validity and reliability. These supporting data are the underpinnings of accurate clinical diagnosis (Hutchinson, 1996; McCauley & Swisher, 1984). This lack of support

is particularly problematic in light of best practice guidelines from ASHA, which indicate that tests with solid reliability and validity be selected for clinical use (ASHA, 2005b). That said, many audiologists still do not know whether the (C)AP tests they use are accurate for diagnosing the existence of a (C)APD (Chermak et al., 2007).

Thus, data from this study suggest that there is work to be done in critically examining tests commonly used to diagnose the presence or absence of (C)APD. First and foremost, more information is needed for potential test administrators with regard to each test's validity (Hutchinson, 1996). Without knowing that a test can be compared to other, similar tests and activities (concurrent and predictive validity), one cannot possibly measure (C)AP skills precisely. Similarly, tests that lack reliability in the form of inter-rater and test-retest reliability are cause for concern, as there is no confidence that test scores recorded at a particular juncture would hold true over time and administrator. Again, these shortfalls pose a tremendous threat to accurately measuring (C)AP capabilities. Thus, there is a great need for additional research to determine the overall efficacy of those assessment tools that lack test validity and reliability data to support or refute their continued clinical application (Keith, 2009a).

Ideally, the first question audiologists should ask in selecting an assessment tool is whether it is able to accurately diagnose the presence or absence of a disorder. Of the 22 tests selected for evaluation in this study, only seven tests (32%) contain acceptable levels of sensitivity for diagnosing (C)APD while eight (36%) report adequate levels of specificity. This indicates that the majority of tests identified as being the most commonly used tests for (C)APD (Chermak et al., 2007; Emanuel, 2002) lack data related to their diagnostic accuracy. Assessment tools that do not have these data reported in examiner's manuals or in refereed journal articles should be used with the utmost caution, as they might well be inappropriate for use in making diagnostic decisions (Spaulding et al., 2006).

If a selected test possesses adequate diagnostic accuracy, audiologists must then use their clinical expertise to carefully consider that assessment tool's overall test validity (Spaulding et al., 2006). Overall, a guiding principle to direct this selection of assessment tools is the notion that if threats to a test's validity are minimal, the test is likely appropriate to consider for clinical use. The converse is also true; that if threats to a test's validity are large in number, then the test is likely inappropriate for diagnostic use. Audiologists need to undertake (C)APD testing with the understanding that no one assessment tool is likely sufficient for use as a basis for diagnostic decision making (ASHA, 2005a; Emanuel, 2002). Rather, a variety of assessment tools need to be used to confirm the presence or absence of (C)APD, and through interpretation of test data, evaluation of each administered test's validity, and through triangulation of all data (including that gathered from non-behavioral tests), a diagnosis likely can be reached (ASHA, 2005a).

Difficulty in standardizing behavioral tests that assess (C)AP abilities in children can be associated with the complexity of separating auditory processing skills from cognitive and language capabilities. To add to this dilemma, various tests used for evaluating (C)AP have been derived from research on adults with identified pathological conditions in the central auditory nervous system (e.g. Dichotic Digits Test, Duration Pattern Test)). However, in children, additional characteristics such as cognition and language can affect the comprehension of auditory information, making it extremely difficult to extricate auditory processing as a discrete entity (ASHA, 2005). Even with the use of electrophysiologic testing, a general form of learning disability may not be delineated from a specific auditory deficit. Obviously, these factors complicate the process of ensuring strong test validity for (C)APD assessment tools. It is imperative, therefore, that clinicians make certain that accurate diagnoses are made with regard to (C)APD, even in the face of such complications. Knowing the validity of behavioral tests commonly administered as part of the (C)APD battery is a step in the right direction.

In the future, specific questions need to be addressed in order to develop a "gold standard" for (C)APD tests. First of all, a determination must be made relative to which criteria should be used for validating sensitivity and specificity of a test in the absence of a normative group with a known neurological lesion. Additionally, guidelines must be developed to explore the relationship between cognition, language skills, and performance on (C)AP tests. Finally, audiologists must determine the best standard for determining when a child falls within the clinical population for a (C)APD in order to make accurate diagnoses. This research will aid with the third charge, but work is needed to begin to address other identified concern.

References

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington DC: author.

American Speech-Language-Hearing Association. (2005a). *Central auditory processing disorders.* Available at http://www.asha.org/members.deskref-journals/deskref/default.

American Speech-Language-Hearing Association. (2005b). *Evidence-based practice in communication disorders* [Position Statement]. Available from http://www.asha.org/policy

American Speech-Language-Hearing Association. (2006). *Preferred practice patterns for the profession of audiology* [Preferred Practice Patterns]. Available from www.asha.org/policy.

Auditec of St. Louis. (n.d.). *Dichotic Digit Test.* St. Louis, MO: Auditec.

Auditec of St. Louis. (n.d.). *Duration Pattern Sequence Test.* St. Louis, MO: Auditec.

Auditec of St. Louis. (n.d.). *Low-Pass Filtered NU-6 Test.* St. Louis, MO: Auditec.

Auditec of St. Louis. (2005). *Masking Level Difference.* St. Louis, MO: Auditec.

Auditec of St. Louis. (n.d.). *Spondaic Binaural Fusion Test.* St. Louis, MO: Auditec.

Baran, J. (2007). Test battery considerations. In: *Handbook of central auditory processing disorder: Volume I: Auditory neuroscience and diagnosis,* G. Chermak & F. Musiek (Eds). San Diego: Plural Publishing.

Bellis, T. (1996). *Assessment and management of central auditory processing disorders in the educational setting: From science to practice*. San Diego: Singular Publishing.

Bellis, T. (2003). *Assessment and management of central auditory processing disorders in the educational setting: From science to practice* (2nd ed.). Clifton Park, NY: Delmar Learning.

Berrick, J., Shubow, G., Schultz, M., Freed, H., Fournier, S., & Hughes, J. (1984). Auditory processing tests for children: Normative and clinical results on the SSW word test. *Journal of Speech and Hearing Disorders, 49*, 318-25.

Bornstein, S., Wilson, R., & Cambron, N. (1994). Low- and high-pass filtered Northwestern University Auditory Test No. 6 for monaural and binaural evaluation. *Journal of the American Academy of Audiology, 5*(4), 259-64.

Brannen, S. (2008). Expand your practice: Add (central) auditory processing services. *Perspectives on Audiology, 4*, 4-8.

Cherry, R. (1998). *Selective Auditory Attention Test*. St. Louis, MO: Auditec.

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.

Emanuel, D. (2002). The auditory processing battery: Survey of common practices. *Journal of the American Academy of Audiology, 13*, 93-117.

Etymotic Research. (2005). *BKB: Speech-in-Noise Test.* Elk Grove Village, IL: Etymotic Research, Inc.

Entwisle, D. & Astone, N. (1994). Some practical guidelines for measuring youth's race/ethnicity and socioeconomic status. *Child Development, 65*, 1521-1540.

Etymotic Research. (2001). *Quick Speech-in-Noise Test* [Audio CD]. Elk Grove Village, IL: Author.

Friberg, J. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*(1), 77-92.

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*, 196-206.

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

Jerger, J. & Musiek, F. (2000). Report of the consensus conference on the diagnosis of auditory processing disorders in school-aged children. *Journal of the American Academy of Audiology, 11*, 467-74.

Katz, J. (1998). *Staggered Spondaic Word Test, 5th edition*. Vancouver, WA: Precision Acoustics.

Katz, J. & Fletcher, C. (1982). *Phonemic Synthesis Test.* Vancouver, WA: Precision Acoustics.

Katz, J., Johnson, C., Brandner, S., Delagrange, T., Ferre, J., King, J., Kossover-Wechter, D., Lucker, J., Medwetsky, L., Richard, S., Rosenberg, G., Stecker, N., & Tillery, K. (2002). Clinical and research concerns regarding the 2000 APD consensus report and recommendations. *Audiology Today, 14,2,* 14-17.

Keith, R. W. (1994). *Auditory Continuous Performance Test*. San Antonio: The Psychological Corporation.

Keith, R. (2000). *Random Gap Detection Test*. Finneytown, OH: Tarton Products.

Keith, R. (2009a). Controversies in standardization of auditory processing tests. In A.Cacace & D. McFarland (Eds). *Controversies in central auditory processing disorder.* (pp. 169-197). San Diego: Plural Publishing.

Keith, R. (2009b). *SCAN-3 for Adolescents and Adults: Tests for Auditory Processing Disorders*. San Antonio, TX: Pearson Education, Inc.

Keith, R. (2009c). *SCAN-3 for Children: Tests for Auditory Processing Disorders*. San Antonio, TX: Pearson Education, Inc.

Keith, R. & McCrosky, R. (1996). *Auditory Fusion Test-Revised*. St. Louis, MO: Auditec.

Kelly, A. (2007). Normative data for behavioral tests of auditory processing for New Zealand school children aged 7 to 12 years. *Australian and New Zealand Journal of Audiology, 29*(1), 60-4.

McCauley, R. (1996). Familiar strangers: Criterion-referenced measures in communication disorders. *Language, Speech, and Hearing Services in the Schools, 27*, 122-31.

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

Merrell, A. & Plante, E. (1997). Norm-referenced test interpretation in the diagnostic process. *Language, Speech, and Hearing Services in Schools, 28*, 50-58.

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

Musiek, F. (1983a). Assessment of central auditory dysfunction: The dichotic digit test revisited. *Ear and Hearing, 4*, 79-83.

Musiek, F. (1983b). *Dichotic Digit Test.* Storrs, CT: Audiology Illustrated.

Musiek, F. (1994). *Duration Pattern Test.* Storrs, CT: Audiology Illustrated.

Musiek, F. (2005). *Gaps-in-Noise Test.* Storrs, CT: Audiology Illustrated.

Musiek, F., Baran, J. & Pinheiro, M. (1990). Duration pattern recognition in normal subjects and patients with cerebral and cochlear lesions. *Audiology, 29*, 304-13.

Musiek, F., Gollegly, K., Kibbe, K., Verkest-Lenz, S. (1991). Proposed screening test for central auditory processing disorders: Follow-up on the dichotic digits test. *The American Journal of Otology, 12*, 109-13.

Musiek, F. & Pinheiro, M. (1987a). Frequency patterns in cochlear, brainstem, and cerebral lesions. *Audiology, 26*, 79-88.

Musiek, F. & Pinheiro, M. (1987b). *Frequency Pattern Test.* Audiology Illustrated: Stors, CT.

Pena, E., Spaulding, T, & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology, 15*, 247-54.

Pinheiro, M. (1977). *Pitch Pattern Sequence Test.* St. Louis, MO: Auditec.

Plante, E. & Vance, R. (1994). Selection of preschool language tests: A data based approach. *Language, Speech, and Hearing Services in Schools, 25*, 15-24.

Plante, E. & Vance, R. (1995). Diagnostic accuracy of two tests of preschool language. *American Journal of Speech-Language Pathology, 4*, 70-76.

Restrepo, M. & Silverman, S. (2001). Validity of the Spanish Preschool Language Scale-3 for use with bilingual children. *American Journal of Speech-Language Pathology, 10*, 382-93.

Riccio, C., Cohen, M., Hynd, G., & Keith, R. (1996). Validity of the Auditory Continuous Performance Test in differentiating central auditory processing disorders with and without ADHD. *Journal of Learning Disabilities, 29*(5), 561-66.

Schow, R., Seikel, A., Brockett, J., & Whitaker, M. (2007). *Multiple auditory processing assessment*. St. Louis, MO: Auditec.

Shea, S. & Raffin, M. (1983). Assessment of electromagnetic characteristics of the Willeford CAP test battery. *Journal of Speech and Hearing Research, 26*, 18-21.

Shinn, J., Chermak, G., & Musiek, F. (2009). GIN (Gaps-In-Noise) permanence in the pediatric population. *Journal of the American Academy of Audiology, 20*, 229-38.

Singer, J., Hurley, R., & Preece, J. (1998). Effectiveness of central auditory processing tests with children. *American Journal of Audiology, 7*, 73-84.

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

Stewart, L. & Kaminski, R. (2002). Best practices in developing local norms for academic problem solving. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 737-752). Bethesda, MD: National Association of School Psychologists.

Thomas-Tate, S., Washington, J., & Edwards, J. (2004). Standardized assessment of phonological awareness in low-income African American first graders. *American Journal of Speech-Language Pathology, 13*, 182-190.

Yavas, M. & Goldstein, B. (1998). Phonological awareness and treatment of bilingual speakers. *American Journal of Speech-Language Pathology, 7*, 49-60.

Willeford, J. (1976). Differential diagnosis of central auditory dysfunction. In L. Bradford (Ed.), *Audiology: An audio journal for continuing education* (Vol. 2). New York: Grune & Stratton.

Willeford, J. (1977). *Competing Sentences.* St. Louis, MO: Auditec.

Wilson, R., Moncrieff, D., Townsend, E., & Pillion, A. (2003). Development of a 500 Hz masking-level difference protocol for clinic use. *Journal of the American Academy of Audiology, 14*, 1-8.